Application of discrete-time survival analysis techniques in modelling student dropout: A case of engineering students at Tshwane University of Technology, South Africa

by

PRINCESS RAMOKOLO

Dissertation

Submitted in fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE

(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

Supervisor : Dr D Maposa
Co-supervisor : Prof M Lesaoana

December, 2020

Declaration

I declare that the thesis hereby submitted to the University of Limpopo, for the degree of Master of Science in Statistics has not previously been submitted by me for a degree at this or any other university; that it is is my work in design and in execution, and that all material contained herein has been duly acknowledged.

Ramokolo, P (B.Sc. Hons. Statistics)
23rd March 2021

PL Ramokolo
Signature

23 March 2021
Date

Dedication

This dissertation is dedicated to my parents, Patricia and Richard Ramokolo, my husband Sipho Masondo and to my kids Nqaba, Pleiades and Sazi Masondo.

Acknowledgements

I am eternally grateful to my Lord and Savior Jesus Christ for the gift of life. I am also grateful for the strength He gave me to persevere until the end. Without Him I would not have made it. To him be the Glory, Honor and Praise now and forevermore.

I also wish to express my sincere gratitude to the following people:

My supervisor Dr Daniel Maposa and co-supervisor Professor 'Maseka Lesaoana for their guidance and assistance throughout my studies.

Tshwane Univeristy of Technology (TUT) for providing data used for this dissertation. Mr Mpho Mveke for the extraction of data from the integrated tertiary system (ITS).

My former line manager at TUT Dr Maupi Letsoalo for encouraging me to register for the degree, for always being available to share ideas and to provide guidance and for believing in me.

Ms Given Luvhimbi from TUT for her support and assistance with study material.

My parents Patricia and Richard Ramokolo for their love, support and for their keen interest in my studies.

My brother Vusa Ramokolo for always being available to look after the kids.

My husband Sipho Masondo for his continuous support and encouragement.

My kids Nqaba, Pleiades and Sazi Masondo for being a source of inspiration.

Abstract

The ever increasing number of students who drop out of university remains a challenge for Higher Education administrators. In response to this, different studies have been conducted globally in order to identify student retention strategies to fix the problem. However, the challenge continues to prevail year in and year out. Most of the studies conducted in South Africa used statistical methods that ignore the temporal nature of the process of student dropout. This study uses discrete-time survival techniques to model the occurrence and timing of undergraduate engineering student dropout at Tshwane University of Technology (TUT). Discrete-time survival analysis techniques allow for a more appropriate utilisation of the longitudinal nature of institutional data, where the time dependence of the data, time-varying factors and time-invariant factors can all be accommodated in the analysis.

The temporal nature of the process of student dropout was analysed for the cohort of students registered in engineering programmes for the first time in 2010 at Tshwane University of Technology using discrete-time survival analysis methods. The cohort was followed for five years from 2010 through 2014, inclusive. Of particular interest was the incidence of dropout, the determinants of dropout, comparison of the single risk discrete-time model with a competing risk discrete-time model, as well as testing for the effects of unobserved heterogeneity. The study used administrative data obtained from the ITS. The logit model was used to estimate the effects of race, gender, Matric performance, performance in Matric Mathematics, residence type, English language status and time on time to dropout with time measured in academic years. A discrete-time competing risk model in the form of a multinomial logit model was also estimated to account for the possible correlation between graduation and dropout. A frailty model assuming a Gaussian distribution for the frailty term was also estimated to account for unobserved heterogeneity.

The study established that the risk of dropout for nonwhite students is significantly higher than that of white students. Furthermore, it was found that the effects of residence type varied with time. For instance, in the first year students with private based accommodation were more likely to dropout compared to those residing on-

campus. On the other hand, in the third year students accommodated in private residences were less likely to dropout than those residing on-campus. The findings also indicate that the effect of having English as a first language as opposed to as a second language on the risk of dropout was only significant in the fourth year such that first language English students were more at risk of dropout compared to second language students. The findings also revealed inconsistencies between the estimates from the single risk and the competing risk model. Moreover, the effect of unobserved heterogeneity was found to be insignificant.

Recommendations from this study are that discrete-time survival analysis model is more efficient than traditional methods used for analysis of student dropout and should therefore be used for analysis of academic outcomes such as dropout. The model can account for the temporal nature of the process of dropout. Both time-varying and time-invariant explanatory variables can be included in the model.The effects of time-invariant explanatory variables that might have time-varying effects can also be investigated.

List of Acronyms

AIC	Akaine Information Criterion
ANCOVA	Analysis of Covariates
APS	Admission Point Score
BIC	Bayesian Information Criterion
CEO	Chief Operating Officer
CHE	Council for Higher Education
DoE	Department of Education
FGS	First Generation Students
HE	Higher Education
HEI	Higher Education Institution
IIA	Independence from Irrelevant Alternative
ITS	Integrated Tertiary System
KM	Kaplan-Meier
LL	Log Likelihood
LR	Likelihood Ratio
MNL	Multinomial Logit Model
OLS	Ordinary Least Squares
PH	Proportional Hazards
UCT	University of Cape Town
UKZN	University of KwaZulu Natal

UoT University of Technology

SAT Scholastic Attitude Score

TU Traditional University

TUT Tshwane University of Technology

Contents

List of Figures

List of Tables

Chapter 1

Orientation of the study

1.1 Introduction and background

This Chapter outlines the background of the study, research problem, the aim of the study, objectives, significance of the study and the outline of the dissertation.

Student retention has, and remains one of the significant areas of discussion in higher education (HE) globally (Berge & Huang, 2004). Improving student retention remains a primary concern for many institutions. Different student retention strategies have been developed to address this challenge; however, the problems and the challenges continue to persist within the system. South Africa, like other countries, continues to experience high dropout rates, and consequently poor retention and graduation rates. At 15%, the graduation rate for HE in South Africa is reported to be one of the lowest internationally (Letseka & Breier, 2008).

Studies on HE performance by undergraduate students in South Africa show that universities of technology (UoTs) (formerly Technikons), tend to experience higher dropouts than traditional universities (TUs) (Letseka & Breier, 2008). Of the 120,000 undergraduate students enrolled for the first time in 2000, 30% dropped out by the end of the year (34% for UoTs and 25% for TUs). Dropouts decreased over the subsequent years, with an overall total of 11% during 2001 (13% for UoTs and 9% for TUs) and an overall total of 9% in 2002 (11% for UoTs and 7% for TUs). Furthermore, for the 2000 cohort of undergraduate students, 58% had dropped out from UoTs at the end of 2004, compared with 38% at TUs (Kraak, 2008).

The annual Council on Higher Education (CHE) VitalStats: Public Higher Education reports for the four-year period ending in 2016 indicate that after six years of enrolment, dropout rates for the 360-credit undergraduate diplomas were persistently high for engineering fields compared to all other fields (CHE, 2014, 2013, 2012, 2011). The dropout rates show a declining trend over the years; however, the numbers continue to

be disturbingly high. A 2011 study based on a 2010 cohort of undergraduate students at the Vaal University of Technology also revealed a high dropout rate for the Faculty of Engineering compared to other faculties (vd Walt & Naidu, 2011).

Dropping out of university presents problems for students, families, educators, administrators and the government. Students leaving university without having completed their studies may be exposed to various psycho-social problems. Examples include: dissatisfaction with university experiences, disruptions of life plans, and being jobless or being engaged in minor jobs to earn much less over a lifetime. Generally, in South Africa, students who do not complete their tertiary qualification will most likely join the millions of unemployed and have no prospects for a decent life (Bokaba & Tewari, 2014). Furthermore, leaving a higher education institution (HEI) without graduating implies a loss in potential earning power and livelihood, lower job prospects, and a weakened ability to accumulate assets and capital, not to mention personal and emotional consequences.

HEIs are also affected by student dropout because much of the funding provided to HEIs is based on student enrolments. The South African government pays universities a subsidy for every student currently enrolled as well as for every graduate, so a failure to complete a degree results in a loss of revenue for an HEI. The loss of revenue is further compounded by the loss of tuition fees as a result of dropout. A report by the then Department of Education (DoE) estimated the cost of the high dropout rates to the country to be R1.3 billion a year (Letseka, 2009).

The persistent challenge of student retention underscores the need for continued research in this area to develop more accurate predictive models of student dropout. This study aims to make a case for using discrete-time survival analysis techniques to model the occurrence and timing of student dropout. A case of engineering students at Tshwane University of Technology (TUT) is considered.

1.2 Problem statement

Student academic outcomes in HEIs have been widely investigated in South Africa. However, most of the studies followed a cross sectional approach and ignored the longitudinal nature of institutional data. In general, most of the models used in these studies can be classified into the following categories: (*i*) logistic regression models (Bokaba & Tewari, 2014; Sartorius & Sartorius, 2013; Zewotir et al., 2011; Eiselen et al., 2007; Lourens & Smit, 2003), (*ii*) data mining models (Kirby & Dempster, 2014; Mashiloane & Mchunu, 2013; Du Plessis & Botha, 2012; Kanakana & Olanrewaju, 2011; Müller et al., 2007; Van der Merwe & De Beer, 2006) and (*iii*) various ordinary least

squares (OLS) models (Van Rooy & Coetzee-Van Rooy, 2015; Van Zyl & Rothmann, 2012). Furthermore, most of the studies focused on important milestones, e.g. graduation, dropout after first year, passing a specific module/subject, etc. and as such the outcome variable was formulated as binary. Under this approach, two-time points are chosen and the probability of success/failure is modelled based on the student's status at the end of the specified time. The timing of the event is not included in the analysis, hence it is not possible to determine when the event occurred. For instance, when the interest is in dropout it is not possible to determine whether the student left the institution in the first semester, the first year, second year or the sixth year, within the two-time points.

Very few studies have analysed the temporal nature of student dropout. Moreover, very few accounted for the competing nature of graduation and dropout. Among the few is Visser & Hanslo (2005), who used descriptive survival techniques to analyse attrition patterns of students at the University of Cape Town (UCT). In one of the first studies using a competing risks approach, Murray (2014) used a continuous-time competing risks approach to identify institutional and student specific factors influencing the type of outcome experienced by undergraduate students when they leave the University of KwaZulu-Natal (UKZN). The number of extra credit points taken (repeated) by a student before leaving the university was used to represent the survival time.

Murray (2014) estimated the cumulative incidence function (CIF) for graduation, voluntary and involuntary dropout from separate cause-specific hazard regression models. The results of the study indicate that when involuntary and voluntary dropouts are treated as competing risks, students with some form of residence-based accommodation graduated more earlier than those without residence-based accommodation. Having some form of financial aid and higher Matric point score also contributed to students to graduate more quickly. On the other hand, when graduation and voluntary dropout are treated as competing risks, the results indicate that having some form of financial aid and residence-based accommodation increased the length of time that students stayed in the system before dropping out involuntarily. Finally, when graduation and academic exclusion are treated as competing risks, access to some form of financial aid and residence-based accommodation assists in preventing students from dropping out voluntarily.

Zewotir et al. (2015) used a competing risk approach to identify factors associated with successful completion or dropout from a master's programme at the UKZN. The study focused on factors associated with the actual number of years it took students to graduate, while treating dropout as a competing risk; and the number of years it took students to drop out, while treating graduation as a competing risk. The multinomial

logit model suggested by Scott & Kennedy (2005) was used to compute the hazard rates associated with dropout and degree completion for each year. The results show that 50% of the master's students had either graduated or dropped out within two years of registration. In terms of financial aid, the results showed that receiving some form of financial funding appears to reduce the length of time it takes a student to drop out from a master's programme. Moreover, receiving some form of financial aid also reduces the length of time it takes to successfully complete a master's programme. When looking at race, the results indicate that race has no significant effect on dropout. However, for the students who eventually graduated, the results show that time to graduation was shorter for African students.

Neethling (2015) also used a multinomial logit competing risk model proposed by Scott & Kennedy (2005) to investigate determinants of both dropout and degree completion at the UCT. Ndlovu (2015) modelled time to graduation at university using different survival analysis techniques. Cox regression model was used as well as its discrete-time extensions. The results showed that in relation to the Cox proportional hazards (PH) model, the degree of flexibility was less as certain variable effects were satisfied to meet the assumption of proportionality by stratifying on those variables, i.e. due to the assumption of proportionality of hazards in Cox regression, the model was fitted with faculty as a stratification variable and the faculty effect was thus sacrificed. On the other hand, the discrete-time model enabled the inclusion of faculty as a covariate and as such the effect of faculty in addition to gender and race could be analysed.

According to Ndlovu (2015), the use of the discrete-time model also provided accrual probabilities from which a graduation profile could be constructed. The results of the test for unobserved heterogeneity indicated that unavailability of variables did not compromise both the Cox regression and the discrete-time model. The study also showed that the discrete-time mixture competing risks model explained graduation better than the cure model. The results on the cure models revealed the presence of a sizable number of students that will eventually not graduate (Ndlovu, 2015).

Weybright, Caldwell, Xie, Wegner & Smith (2017) used discrete-time Cox regression model to analyse the relationship between substance abuse, leisure experience and school dropout, controlling for demographic and educational factors that have been found to be significantly related to dropout. Besides the work done by Ndlovu (2015) where time to graduation at university was modelled using different survival analysis techniques including discrete-time mixture model to account for the presence of unobserved heterogeneity, there is no record of studies that focused on academic outcomes, specifically student dropout from a competing risks approach utilising techniques that account for unobserved heterogeneity and/or dependent risks.

Ishitani (2008) relates the advantages of applying discrete-time survival analysis in studying student dropout over other techniques. Discrete-time survival analysis is optimal in that enrolment status information of students from different time points (including censored observations) can be incorporated; it is suited to estimating the probability of dropout at different time points; it allows for the examination of probability of highly skewed binary dependent variables; and both time invariant and time-varying covariates can be easily incorporated into the model. According to Kim (2014); Singer & Willett (2003); Yamaguchi (1991), discrete-time survival analysis techniques are appropriate for tied events such as data recorded in retention studies (month, term, semester, year), as they can be handled in an unbiased manner. In essence the discrete-time survival analysis provides an ideal framework, not only for answering descriptive questions around student dropout, but for modelling the relationship between dropout and its predictors as well (Singer & Willett, 1993).

The effects of failure to control for unobserved heterogeneity have been studied. It has been shown that failure to control for unobserved heterogeneity produces biased estimates in single risk models (Van den Berg, 2001; Lancaster, 1985; Heckman & Singer, 1984a; Lancaster, 1979) as well as competing risks survival analysis models (Van den Berg, 2001; Butler et al., 1989). Furthermore, it has been shown that failure to control for unobserved heterogeneity results in negative duration dependence such that increasing event hazard probabilities were diminished over time and decreasing hazard probabilities were accelerated over time (Keiding et al., 1997; Hougaard, 1995; Trussell & Richards, 1985; Heckman & Singer, 1984b). In the current study, it is not possible to account for all variables that have been shown to affect dropout. The study is limited to the variables that are available on the university database, hence the need to control for unobserved heterogeneity.

1.3 Aim of the Study

The purpose of this study is to analyse the temporal nature of process of engineering student dropout at TUT using discrete-time survival analysis methods.

1.3.1 Study objectives

The main objectives of the study are to:

- identify determinants of dropout,
- analyse the incidence of dropout,
- compare the risk profile of dropping out among different groups of students,

- compare the discrete-time single-risk model with the competing-risk model,
- test the effects of unobserved heterogeneity in the single-risk model.

The data used in the study is extracted from the TUT Management Information System. These data cover all first year students enrolled for undergraduate programmes in the Faculty of Engineering and the Built Environment at TUT Pretoria West campus, Gauteng, at the start of the 2010 academic year. These are all three-year programmes which should be completed within three years. Students are tracked for a period of five years. According to Blom (2014), tracking cohorts through undergraduate study in South Africa for the purpose of estimating completion or dropout rates requires a minimum time frame of four years. The student data from January 2010 until December 2014 was considered.

The data contains detailed transcript data from first year to dropout or graduation, personal characteristics, measures of their Matric performance and information on type of accommodation at each point in time. Dropout will be inferred from enrolment records. For students who are still enrolled and have not achieved the diploma by the end of the observation period, the duration is marked as censored. Similarly, students who graduate within the five years are considered censored since they would not have experienced the outcome of interest.

1.4 Significance of the study

This study aims to improve on the current body of knowledge by analysing the temporal nature of student dropout using discrete-time survival analysis methods that account for the effects of unobserved heterogeneity. It further attempts to highlight additional knowledge gained by including the duration of survival, time-invariant factors and time-varying factors as a basis for understanding the underlying pattern of student dropout. Time to dropout is initially estimated as a single event, then the robustness of the results is tested by estimating a competing risk model with dropout and graduation jointly estimated. By considering both the discrete-time single risk and competing risk models, the study allows for investigation of the competing nature between dropout and graduation. The single risk model is further modelled accounting for unobserved heterogeneity. This allows for investigation of the effects of unobserved or unmeasured factors.

The discrete-time single risk as well as the competing risk model for dropout proposed in this study will help in understanding factors that influence dropout of undergraduate engineering students at TUT. The models can be generalised to assist in early iden-

tification of undergraduate engineering students who are more likely to be at risk of dropping out at TUT and other institutions similar to TUT. Furthermore, the models can be used to determine when undergraduate engineering students at TUT and other similar institutions are at the greatest risk of dropping out. Knowing when students are more likely to be at risk of dropout will assist administrators to develop appropriate intervention strategies and remedial programs at the identified risk periods, considering the risk profile. The interventions can be proactive in a prioritised manner taking into account the limited academic resources.

1.5 Dissertation outline

This dissertation is organised into five chapters. Chapter 2 presents a comprehensive literature review of literature related to student dropout in HE. Chapter 2 also introduces survival analysis and its applications in HE student academic outcomes. Special attention is paid to the single risk discrete-time model as well as the competing risk discrete-time models. The Chapter also includes summary findings on factors associated with HE academic student outcomes in South Africa. In Chapter 3, the maximum likelihood estimators for the single risk discrete-time model as well as the discrete-time competing risk model are presented. The Chapter begins with a presentation of continuous-time survival analysis functions as a basis for the discrete-time case. A model that accounts for unobserved heterogeneity in the single risk case, model assessment and diagnostic methods are also outlined. The results of the data analysis based on the models outlined in Chapter 3 are presented in Chapter 4. A description of the data used in the study is also presented in Chapter 4. The study closes with a discussion of the results, conclusions and recommendations in Chapter 5. The limitations of the study are also presented in Chapter 5.

Chapter 2

Literature Review

2.1 Introduction

This chapter presents a literature review of the analysis of the process of student dropout. The chapter starts with a brief review of the definition of dropout along with a discussion on early theories of the process of student dropout. Techniques used in previous studies to analyse dropout as well as their advantages and limitations are also discussed. Emphasis is placed on the origins and development of survival analysis techniques up to its application in student academic outcomes, and in particular dropout. The chapter concludes with a concise review of factors associated with student academic outcomes in South Africa.

2.2 Student retention/dropout

Both nationally and internationally, there is no standardised terminology used to describe and measure student success or academic outcomes across institutions, thus making the task of explaining and understanding student academic success a complex exercise (Letseka et al., 2010). Nevertheless, most definitions embrace the idea of persistence to the completion of the student's enrolled program. According to Hagedorn (2005), student retention and dropout have been the most common measures used in educational research to measure academic student outcomes. Increased retention is the focus of many institutions' quality measures and improvement efforts (Berge & Huang, 2004).

Student retention/dropout can be measured from multiple viewpoints: namely, institutional, system, academic discipline and by course (Hagedorn, 2005). Institutional retention is based on the percentage of students who return to the same institution year after year until the completion of their studies. This is the most commonly used

method by universities and colleges to measure their performance (Ashby, 2004). System retention involves tracking students across all HEIs instead of only the institution they are enrolled in. This means that a student who leaves a particular institution, and enrols at another institution where he/she completes his/her studies, is considered retained within the system of HE. This measure requires tracking students at a national or even international level, which can be costly and difficult to implement. When it comes to academic discipline retention, the focus is on retention within a specific academic discipline. From this viewpoint, students who enrol in a HEI with a statistics major and later change to another major, are classified as not retained.

Retention at an individual course level focuses on classes with low levels of student retention within an institution. Measuring retention from this angle poses complications since one must decide on the number of class sessions sufficient to constitute retention (Styron Jr, 2010). In general, these viewpoints can sometimes contradict each other. For example, students who change institutions may still graduate, however, their departure from the initial institution is in opposition to their goal of retaining its students until graduation. On the other hand, their departure does not adversely affect the overall HE system retention goals. Researchers need to be careful to choose an operational definition of retention that is suitable for the research problem under investigation (Bean & Metzner, 1985). Given the purpose of the study, the goal is that of a single institution and student retention is associated with membership at a specific institution, rather than membership in HEIs in general.

The study focuses on dropout within a specific academic discipline at an institutional level. The definition provided by Fowler & Luna (2009) and Hagedorn (2005) is adopted in the study. The authors define retention in HE as students' continued enrolment until successful completion of an enrolled program (Fowler & Luna, 2009; Hagedorn, 2005). On the other hand, dropout is typically defined as the opposite of retention (Hagedorn, 2005). According to Pocock (2012), dropout is the common terminology used in the literature to describe students who leave a specific HEI without completing a qualification in their chosen initial degree. This definition has been criticised by Astin (1975) for being simplistic and for incorrectly defining dropout, as the so-called dropouts may eventually become non-dropouts by returning to the institution and vice versa. However, the author concedes that there seems to be no practical way out of the dilemma as an accurate classification of dropouts versus non-dropouts can only be attained when all the students have either died without graduating or have graduated.

In order to address this shortcoming, Tinto (1993) distinguished stopout from dropout. Tinto (1993) defined stopouts as students who, after leaving an HEI, return at a later time to complete their degrees, while dropouts are students who leave an HEI and do

not return. It should be noted that studies on student dropout tend to focus on specific predetermined periods. Furthermore, it is not possible to follow students indefinitely. Consequently, from an operational point of view, the definitions of dropout and stopout are guided by the study reference period.

Most students enrolled at TUT come from disadvantaged backgrounds and rely on bursaries, national student financial aid scheme (NFSAS), and other student loans. It is, therefore, reasonable to assume that most of the dropouts are due to poor academic performance resulting in exclusions and loss of funding. The decision to dropout is, therefore, imposed on the students and hence not voluntary. For the purpose of this study, no distinction is made between voluntary and involuntary dropout. It is also important to note that since most students at TUT come from poor communities, it is reasonable to assume that once they dropout, they have a very small chance of re-registering for their studies. Dropout is, therefore, treated as permanent. Any student who ceases to enrol without having attained a diploma is regarded as a dropout. The outcome variable in the study is inherently dichotomous. Students either dropout from an engineering major or persist and graduate. Consequently, the focus is on analysis and prediction of a dichotomous dependent variable.

2.3 Studies on student dropout

Student dropout/retention is one of the most extensively studied topic in HE (Tinto, 2010). Most of the early studies in this area were descriptive in nature (Bean, 1980; Tinto, 1975). Vincent Tinto was one of the early scholars to introduce student retention models in response to the shortcomings associated with the descriptive nature of prior research (Tinto, 1975). He advanced the work by Spady (1970) and formulated the Student Integration Model to describe the process of interaction between an individual student and the institution resulting in dropout by different students from colleges. Tinto (1975) defined student retention as a longitudinal process where the decision to drop out or persist is being influenced by the students' pre-enrolment characteristics, background variables and commitment levels, which are then moderated by their social and academic integration into the institution.

Tinto's theory has since emerged as the most significant theoretical viewpoint among the many theories and frameworks established to describe the process of student dropout (Aljohani, 2016). The model has been tested and adopted in different HE institutions and environments resulting in more credibility and validity (Aljohani, 2016). Tinto's theory has become a basis for many other theories and models, e.g. Bean's (1980) Student Attrition Model; Pascarella & Terenzini's (1980) Student-Faculty In-

formal Contact Model; Bean & Metzner's (1985) Non-traditional Undergraduate Student Attrition Model; and Cabrera, Nora & Castaneda's (1993) Student Retention Integrated Model. When it comes to data analysis, Tinto recommended the use of longitudinal path analysis to analyse the student dropout process (Pantages & Creedon, 1978). Most of the other theories were tested and validated through structural equation modelling, path analysis and regression analysis (Hiemstra, Otten & Engels, 2012; Ishitani & DesJardins, 2002).

The vast body of research on student retention theory has been beneficial in providing a starting point in terms of investigating independent variables that can be included in student retention, dropout model specification; as well as in identifying possible data sources (Bogard, Helbig, Huff & James, 2011). Furthermore, the theoretical models developed have provided useful tools for development of academic and student affairs based retention intervention services (Kerby, 2015). However, these studies do not provide the much-needed instrument to accurately predict retention/dropout (Delen, 2010). The ability to accurately predict student dropout behaviour to develop preventative measures is more important than understanding the reasons behind the behaviour.

Predictive modelling is a frequently used method for which a model is developed to best predict the probability of an outcome of interest (Geisser, 1993). A predictive model quantifies the likelihood that an observation within a sample or population experiences the event or outcome of interest. A probability is attached to each observed unit of the population or sample. Two approaches are used in statistical modelling. The first approach assumes that data are generated by a known stochastic process and the other approach uses algorithmic models and treats the data mechanism as unknown (Breiman, 2001).

According to Ingersoll, Lee & Peng (2010) and Nisbet, Elder & Miner (2009), classical or stochastic studies use past information to determine a future state of a system (often called prediction), whereas the algorithmic approach, also called data mining, uses past information to construct patterns based not only on input data, but also on the logical consequences of those data. This process is also called prediction, however, it contains an element missing in classical techniques, i.e. the ability to provide an orderly expression of what might be in the future, compared to what was in the past (Nisbet et al., 2009). Classical methods include techniques such as linear regression, discriminant function analysis, logistic regression, and analysis of variance. The algorithmic or data mining paradigm includes techniques such as neural networks and decision trees.

Logistic regression has been the most used classical method for analysis of student dropout due to the binary nature of student dropout. It is an appropriate analytical

tool when the interest is to describe and test hypothesis about the relationship between a binary or dichotomous categorical outcome variable and one or more continuous or categorical predictor variables (Hosmer & Lemeshow, 2000). Logistic regression has been shown to be superior to its alternative counterparts, e.g. discriminant function analysis, log-linear models and linear probability models, mainly based on the fact that (i) it can accept both continuous and discrete explanatory variables, (ii) it is not restricted by normality or equal variance/ covariance assumptions for the residuals, and (iii) it is related to the discriminant function analysis through the Bayes theorem (Ingersoll et al., 2010; Peng, Stage, St John & So, 2002). According to Dey & Astin (1993), violations of these assumptions can lead to biased estimates. Furthermore, linear models were also found to sometimes predict values for dichotomous variables that have no meaning such as negative probabilities or probabilities that exceed 1 (Pohlman & Leitner, 2003; Agresti, 1990).

The other concern is whether the relationship can be truly linear when dealing with probabilities as changes in the independent variable are likely to have more impact on the probability of an event occurring at the middle of the probability range than at the end of it (Agresti, 1990). The use of logistic regression for analysis of binary outcome variables in higher education can be traced to the late '60s and early '70s (Cabrera, 1994). As early as 1975, Tinto called for the use of logistic regression to study student college retention because of the categorical nature of dropout as an outcome variable (Dey & Astin, 1993). Examples of studies that used logistic regression to analyse student dropout include: Rohr (2012), Wohlgemuth et al. (2007), Pyke & Sheridan (1993), Cabrera, Stampen & Hansen (1990), Neumann & Finaly-Neumann (1989) and Stampen & Cabrera (1986).

Data mining has been another technique used to predict student dropout in HE. Data mining refers to the extraction or "mining" of knowledge from large amounts of data (Han, Kamber & Pei, 2006). The application of data mining methods and tools for analysing data available at institutions of higher learning defined as Educational Data Mining (EDM), is an emerging new stream of data mining research (Kabakchieva, 2013; Romero & Ventura, 2013). Studies focussing on dropout tend to invoke classification tasks due to the categorical nature of dropout. Different classification algorithms have been used to analyse student dropout, e.g. decision trees (Jadrić, Garača & Čukušić, 2010; Yu, DiGangi, Jannasch-Pennell & Kaprolet, 2010; Dekker, Pechenizkiy & Vleeshouwers, 2009; Al-Radaideh, Al-Shawakfa & Al-Najjar, 2006), Logistic regression (Dekker et al., 2009; Jadrić et al., 2010), Bayesian classifier (Dekker et al. (2009); Kotsiantis, Pierrakeas & Pintelas (2003)), random forest algorithm (Dekker et al., 2009); and neural networks (Tan & Shao, 2015; Delen, 2011; Jadrić et al., 2010; Yu et al., 2010).

2.4 Shortcomings of previous studies

Most of the studies in the literature used longitudinal data to estimate the probability of dropout, however, they mostly only looked at two time-points. A set of relevant explanatory variables associated with dropout is collected at a chosen initial time point and then collected again after a specified period. The effect of these covariates on dropout is then estimated. This is done for the general student population and for specific groups. The timing of dropout is not included in the analysis; and hence these studies cannot respond to the question of when these dropouts occur (Willett & Singer, 1991). Furthermore, the observations are separated into those that have experienced the event and those that have not experienced the event within the specified time.

This dichotomised sample can hide knowledge about educational transitions (Willett & Singer, 1991). It could also potentially remove meaningful differences in event times by grouping together everyone who has experienced the event and those who have not by the chosen cut-off time (Willett & Singer, 1991). For example, when the outcome of interest is graduation, students who graduate in three years are not distinguished from those who graduate in four years. Using combined periods of time also means that the results could vary depending on the different time periods combined or the end points of the research studies (Willett & Singer, 1991).

According to Herrera (2006), many variables vary in their success at predicting dropout, depending on academic level, i.e. variables which have a significant effect on dropout at one academic level will not necessarily have the same effect at a different academic level. The time-varying nature of some key predictors/explanatory variables, the non-constant level of risk that individuals may experience over time, and various lengths of exposure to risk among the participants, cannot be addressed by models with one record (or observation) per participant (Ampaw & Jaeger, 2012). Time-varying variables are modelled as different variables giving up degrees of freedom in estimation (Ampaw & Jaeger, 2012). Consequently, these studies produce static models. The process of student dropout is assumed to be uninform and the dynamic nature of the student dropout process is ignored (DesJardins, Ahlburg & McCall, 1999). This, for instance, means that there is no difference in the dropout behaviour of first year students and senior students.

Logistic regression, structural equation modelling (SEM) and data mining cannot correctly account for students who have not dropped out by the end of the research period. It is assumed that these students will never dropout. In general, the methods cannot account for observations that have not experienced the event of interest by the end of the study period. This underestimates the probability of dropout and leads to biased

estimates (Ameri, Fard, Chinnam & Reddy, 2016). In general, these methods are not suited for analysis of time to event occurrence as they do not account for censoring and the changes in the risk of dropout over time, and they do not provide details about the risk of dropout over time.

2.4.1 Censored observations

Willett & Singer (1988) highlighted the challenges of building models of time to event occurrence as a function of selected predictor variables. In particular, analysis of time to event dropout is complicated by incomplete data with regards to the value on the dependent variable, e.g. time to dropout. This means that sometimes it is only known that the survival time T is larger than some censoring time C. For example, some students might not leave university during the observation period. For these students, it would not be possible to determine whether they dropped out or when they had dropped out. This kind of missing data is referred to as right censoring, also referred to as Type I censoring (Allison, 2010). This means that the censoring is fixed (under the control of the investigator) and all the observations have the same censoring time (Allison, 2010).

Right censoring can also occur when a participant in a study withdraws prematurely. For example, in trials of a new drug therapy, a patient might experience severe side effects and therefore stop participating. Participants might also be lost to follow up, i.e. they might disappear for unknown reasons as in the case of longitudinal studies where some participants relocate and cannot be traced. This is referred to as random right censoring, and it is similar to Type I censoring, except that the time of censoring is itself a random variable, that is, it is not a fixed or pre-specified value (Allison, 2010). The other type of right censoring is Type II censoring, where a sample is observed until a pre-specified number of events has occurred (Allison, 2010). A common study design that results in Type II censoring involves animal experiments where the study is stopped after k deaths occur, where k is determined to be the minimum number of event times needed for sufficient statistical power.

Censoring can also be interval based, such that a participant is only known to have experienced the event of interest within two-time periods, but the exact time is unknown (Cleves, 2008). In practice, this occurs when participants are evaluated or examined at fixed time points throughout the follow-up period. Interval censoring is common in clinical trials and longitudinal studies with regularly timed follow-up assessments. For example, some clinical outcomes of interest can only be determined by a physician's examination: at one assessment, the subject is considered disease-free and at the next assessment, the subject is diagnosed as having the disease. In such cases, it can be

difficult, if not impossible, to determine when exactly between the two assessments the subject developed the disease. Discretely measured survival data can be considered a special case of interval censored data, when all participants that experience the event during the observation period are interval censored and the possible intervals of censoring are common to all participants, e.g., all participants are assessed at the same follow-up times. Both Type I and random right censoring can be reformulated as special cases of interval censoring, with left censored individuals experiencing the event in the interval from zero to the time of first observation; and right censored individuals experiencing the event in the interval from the time of last observation to infinity.

Observations can also be left censored, i.e. when a participant in the sample has experienced the event of interest prior to the onset of observation (Hosmer & Lemeshow, 2000). In this case, all that is known about the event timing is that the survival time T is smaller than C, where C denotes the time until the end of study or loss to follow-up. For example, this can occur in a study where the event of interest is the age at which a child learns to accomplish certain tasks in children learning centres. Left censoring occurs if children can already perform the tasks when they start their study at the centres. Mathematically, left censoring is also not different from interval censoring (Cleves, 2008). In both cases, the event occurred at some time when the participant was not under observation, in this case, prior to the participant being observed, and hence happened in an interval.

2.4.2 Handling censored observations

The presence of censored observations renders classical statistical techniques inappropriate for analysis of time to event studies (Lee & Go, 1997). Different strategies have been employed to address the challenge posed by censoring. Some studies used classical statistical techniques like OLS and focused only on observations with uncensored event times (Baird, 1990; Abedi & Benkin, 1987). However, according to Allison (1982), in the presence of censored observations, the analysis of data based only on observations with uncensored event times, results in underestimation of the actual time to event occurrence. For instance, in a study focusing on time to completion of a doctoral degree qualification, the median lifetime computed based on only students who have completed the degree, will be less than the true median due to exclusion of students who have not yet completed their degrees. Exclusion of censored observations alters the distribution of time to graduation. The very presence of ungraduated students is an indication that the true time to graduation is much longer than the one estimated based on only students who have completed their degrees. In general, the presence of censored observations is in itself informative. It implies that time to event occurrence

is actually longer that the specified time period. Ignoring the censored observations can result in loss of efficiency due to a loss in sample size (Leung, Elashoff & Afifi, 1997) and severe bias (Allison, 2014).

The other option is to impute the missing survival times. According to Leung et al. (1997), two approaches can be employed in this regard. Firstly, the censored observations can be assigned the value of the study endpoint time. Singer & Willett (1993) explain that this kind of imputation changes non-events into events and further assumes that all these new events occur at the earliest times possible. Secondly, it can be assumed that all censored observations will never experience failure. The two approaches result in overestimation and underestimation of the survival probabilities respectively, rendering them both inappropriate.

Another approach assumes that time to event after censoring follows a specific model for which parameters are estimated to impute the residual time from censoring to event occurrence (Leung et al., 1997). This approach has been criticised for relying too heavily on the model assumptions, which are difficult to verify without information on time to event after censoring (Leung et al., 1997). The challenge of censoring can also be circumvented by focusing only on event occurrence versus non-occurrence within a fixed period and ignore survival times. This approach results in dichotomised data hence suffers the same shortcomings as logistic regression.

2.5 Survival analysis

Survival analysis has been proposed as a distinctive and effective technique for analysis of HE research data that has an emphasis on longitudinal student outcomes such as student dropout (Yamaguchi, 1991). It is a branch of statistics that involves modelling time to event data whilst handling the challenges posed by censoring even-handedly. It is one of the oldest fields of statistics, going back to the 17th century. The motivation for this method was initiated in the analysis of clinical trials data with time to death as the outcome of interest, hence the term "survival analysis" (Fleming & Lin, 2000). While the terminology survival analysis has commonly been used in the medical field, in recent years it has come to be known by different names in different areas of study: event history (sociology); reliability analysis (engineering); failure time analysis (engineering); duration analysis (economics); transition analysis (economics); and survival analysis (medicine) (Allison, 1982). It has increasingly been applied in many other fields. For instance, it has been used in socio-economic studies to investigate issues such as employment/unemployment, inflation, tourism demand, tropical deforestation and many others (Lee, Lee & Kim, 2017; Barros, Butler & Correia, 2010; Leung, Rigby

& Young, 2003; Vance & Geoghegan, 2002; Narendranathan & Stewart, 1993; Kiefer, 1988). In finance and the banking industry some of its applications include credit model development, assessment of possible exit options and the timing of exit of venture capitalists, investigation of time series dependence in the direction of stock prices and analysis of hedge funds and commodity trading advisors (Marimo & Chimedza, 2017; Ju, Jeon & Sohn, 2015; Giot & Schwienbacher, 2007; Lunde & Timmermann, 2004; Stepanova & Thomas, 2002; Brown, Goetzmann & Park, 2001).

Li (2014) used survival analysis in industrial engineering to estimate and predict different duration stages of traffic incidents occurring on urban expressways. Survival analysis has also been used in engineering to predict life expectation and product reliability (Pham, Yang & Nguyen, 2012; Meeker, Escobar & Hong, 2009; Mueller et al., 2007; Hough, Garitta & Gómez, 2006). In marketing, it has been used to analyse customer behaviour such as adoption of new products (Bilgicer, Jedidi, Lehmann & Neslin, 2015), customer life-time duration (Meyer-Waarden, 2007) and occurrence and timing of repeat purchase (Ansell, Harrison & Archibald, 2007; Harrison & Ansell, 2002). In sports, survival analysis has been used to model recurrent sports injuries (Ullah, Gabbett & Finch, 2014); to model time to first and second goal occurrence in soccer (Nevo & Ritov, 2013); to compare time to death of Olympic medallist to the general population (Clarke et al., 2012) and to investigate transitions into and out of regular sports and exercise participation (Lunn, 2010).

The use of longitudinal data in survival analysis makes it possible to determine at what time periods the event of interest is most likely to occur, as well as to determine why some individuals experience the event earlier than others, and to also determine why some individuals do not experience the event of interest at all during the study period (Min, Zhang, Long, Anderson & Ohland, 2011; Murtaugh, Burns & Schuster, 1999). Survival analysis makes efficient use of data available from all subjects: those who experience the event of interest and those who do not during the study period (censored), and as such, overcomes the difficulty of handling censored observations (Min et al., 2011; Lesik, 2007; Jonson, 2006). With this approach, the outcome of interest in retention studies can be reframed from whether an event occurs to when does it occur. This permits a more appropriate utilisation of longitudinal data and sample attrition problem encountered in survey studies can be avoided. In retention studies, longitudinal data analysis enables researchers to follow factors that impact a student's decision to stay or drop out of their studies over a period of the study (Willett & Singer, 1991), and hence allows for the use of the most recent information in the analysis.

Although survival analysis is one of the oldest fields of statistics, going back to the 17th

century, its application in student retention studies was only introduced in the late 80s. In one of the early applications of survival analysis in the field of education, Willett & Singer (1988) used a data-based example from an investigation of ten-year long teacher survival patterns for a cohort of teachers who started their profession in 1972 in Michigan to establish a framework for doing good data analysis with proportional hazards models. In 1991, the authors used the same data set to show what could be learned about educational transitions by answering the question of whether events occur by trying to determine when the events occurred (Willett & Singer, 1991). They also suggested that researchers should examine when an individual was at the greatest risk of experiencing the event of interest. For instance, rather than asking whether students drop out before a certain time, they suggested that researchers should focus on when they are at the greatest risk of dropping out (Willett & Singer, 1991). They used survival analysis to describe teachers' career transitions by building statistical models of risk of event occurrence over time (Willett & Singer, 1991).

Murtaugh et al. (1999) followed in Willett & Singer (1991) footsteps. They demonstrated the advantages of using survival analysis to analyse student retention data. Their study also determined some of the factors that affected student retention at Oregon State University. Their results showed that dropout tends to occur in tempos at the end of each school year with swift declines at the end of the students' first spring quarter. DesJardins et al. (1999) also used survival analysis to better investigate the process of student dropout from university. The models with only time-invariant covariates were compared to those with time-varying covariates. The results of their study indicated that the effects of covariate on the probability of the first stopout varies with time. The addition of timing into the model improved the understanding of dropout.

Ishitani & DesJardins (2002) investigated time to dropout over a period of five years. Their study showed the time-varying effects of factors associated with student dropout. For instance, both the time-invariant and the time-varying models showed that students from lower income families were more likely to dropout than those from more affluent families, after accounting for other variables. However, this relationship varied over time, i.e. the estimated coefficients for students from low income families were 0.73 in the first year, 0.95 in the second year, 1.07 in the third year and 0.78 in the fourth and fifth years. Among other factors, the results also showed that the effects of scholastic attitude test (SAT) scores on dropout varied with time, i.e. in the first year, the risk of dropout for students with SAT scores in the highest quartile was 35% lower than that of those in the lowest SAT quartile. In the second year, the effect of highest quartile SAT was not significant. The use of a survival analysis models to analyse the process of student dropout is now well established (Ameri et al., 2016; Alarcon & Edwards, 2013; Yang et al., 2013; Chen & DesJardins, 2010; Murphy et al., 2010; Bruinsma &

Jansen, 2009; Lesik, 2007; DesJardins et al., 2002; Murtaugh et al., 1999).

Three features must be correctly defined in survival analysis, i.e. the target event, starting point, and the time metric. It must be clear what represents the target event of interest and which transition between states is of interest. The beginning of time is the point at which all observations in the study are at risk of experiencing the event of interest. For example, students become at risk of dropping out the date at which they enrol at university. An event occurs when observations move from their current state to another. The distance between the beginning of time and event occurrence is called event time enrolment at university. The time metric in which event time is recorded needs to also be clearly specified. Time can be recorded in a fine-grained time metric (thin precise units), e.g. seconds, minutes, hours, days. Examples of such events are death and injury.

Event times can also be measured in discrete-time points, e.g. months, semester, and year. Some events occur at truly discrete-time points, while others occur at continuous-time points, but are recorded in discrete-time points (grouped continuous time observations) (Scheike & Jensen, 1997). For example, students doing semester courses complete their studies only one time per semester such that their data is only available in discrete-time points, which is at the end of the semester. On the other hand, a student may drop out at any point during the semester, but that data is only available at the end of the semester. The event is consequently recorded in discrete-time points even though the timing is continuous.

Different survival analysis methods are used depending on the metric for time. Continuous-time survival analysis techniques are used to analyse survival data recorded on a continuous scale. In terms of discrete-time survival data, discrete-time survival analysis methods are recommended for the case where events occur at truly discrete-time points (Allison, 1982). In the second case where events occur at continuous-time points, but are recorded in discrete-time points, the time metric can be treated as continuous and continuous-time survival analysis techniques can be used. Alternatively, discrete-time survival techniques can be used. Given that both approaches do not affect model specification and the advantages of using discrete-time techniques, discrete-time models are recommended (Allison, 1982).

Ishitani (2008) relates the advantages of applying discrete-time survival analysis in studying student dropout over other techniques. Discrete-time survival analysis is optimal in that enrolment status information of students from different time points (including censored observations) can be incorporated. It is suited to estimating the probability of dropout at different time points. It allows for the examination of probability of highly skewed binary dependent variables, and both time-invariant and time-varying

covariates can be easily incorporated into the model. According to Kim (2014), Singer & Willett (2003) and Yamaguchi (1991), discrete-time survival analysis techniques are appropriate for tied events such as dropout, as they can be handled in an unbiased manner. Discrete-time models are also preferred over continuous-time models since the magnitude of the baseline hazard rate cannot be estimated through continuous-time models (Chen & DesJardins, 2010). Additionally, discrete-time models do not require the proportionality of hazard assumption in the presence of time-varying covariates (Muthen & Masyn, 2005).

Most survival analysis studies on student academic outcomes followed a discrete-time approach, e.g. dropout (Min et al., 2011; Gury, 2011; Chen & DesJardins, 2008; Lesik, 2007; DesJardins et al., 1999; Murtaugh et al., 1999; Willett & Singer, 1991), and graduation (Ampaw & Jaeger, 2012; Aina et al., 2011; Murphy et al., 2010; Bruinsma & Jansen, 2009; Doyle, 2009). However, some studies treated time as continuous. For instance, Chimka et al. (2007) investigated engineering college student graduation patterns using the Cox Proportional Hazards (PH) model. Time to graduation was treated as continuous and the Breslow (1974) method for handling tied survival times was used in the estimation of the models. Restaino (2008) analysed the interval between the first enrolment at university and the first occurrence of non-enrolment. In his study, the covariates were assumed to be time-invariant, time to dropout was assumed to be continuous and it was also assumed that there are no ties among the dropout times.

The choice between parametric, non-parametric and semi-parametric techniques is also important. In most cases, the exact distribution of event times is usually unknown. Non-parametric models are more reliable is such cases as they have greater flexibility and protect from the dangers of misspecification. However, the methods are essentially descriptive, and can only be used when the model involves no covariates. On the other hand, semi-parametric models allow for inclusion of covariates without making any assumption about the baseline hazard function, but only assumes parametric form for the effect of the explanatory variables on the hazard.

Non-parametric and semi-parametric models are the most widely used survival analysis techniques. This is also true in student retention studies as evidenced by the many studies using these approaches compared to parametric techniques (Ameri et al., 2016; Paura & Arhipova, 2014; Min et al., 2011; Gury, 2011; Bowers, 2010; Nicholls et al., 2010; Reibnegger et al., 2010; Plank et al., 2008; DesJardins, 2003; DesJardins et al., 1999). Radcliffe et al. (2006) and Ishitani (2003) are among the few researchers that followed a parametric approach when analysing student retention.

2.5.1 Discrete-time single risk models

The hazard rate is the ultimate dependent variable in survival analysis (Allison, 1984). The sample hazard function and the survival function provide a good summary of the estimated population profile of risk, and they allow us to see whether and when an event is likely to occur. The timing of dropout can be modelled by estimating the risk or hazard rate of dropout in each semester over the observation period (Singer & Willett, 2003). Generalised linear models are used as the basis of analysis of hazard probabilities as a function of covariates for discrete-time survival data. In generalised linear models, the probability for a categorical outcome is transformed by a link function and modelled as linear with respect to explanatory variables. The use of an appropriate link function ensures that the estimated probability lies between 0 and 1. The logit and the probit link functions are the most commonly used link functions for categorical outcomes (Box-Steffensmeier & Jones, 2004). However, between the two, the logit link function is more popular. Its use was proposed by Cox (1972) as the discrete-time model counterpart to his continuous-time regression model.

The main reason for the popularity of the logit link function over other alternatives is that the full maximum likelihood estimates for the parameters of the discrete-logit model can be estimated using the regular logistic machinery available in most statistical analysis software (Box-Steffensmeier & Jones, 2004). It has been used, for example, in transportation studies to model the time that transpires until a trip is taken (Hensher & Mannering, 1994); in mental health to analyse patterns of remission from substance use disorder (Xie et al., 2003); in policy studies to analyse adoption of policy (Jones & Branton, 2005); in veterinary medicine to investigate risk factors associated with fatal injuries of animals (Henley et al., 2006); in banking to analyse risk of default (De Leonardis & Rocci, 2008) and to predict bankruptcy(Nam, Kim, Park & Lee, 2008); in substance abuse studies to examine the timing of smoking onset during mid or late adolescence (Hiemstra et al., 2012); to identify factors associated with initiation to inhalant use among adolescents (Nonnemaker, Crankshaw, Shive, Hussin & Farrelly, 2011); and to investigate the effect of truancy on the initiation of marijuana use (Henry et al., 2009), in entrepreneurship to measure the risk of closing business and probability of businesses opening (Yoon & Currid-Halkett, 2015); and to analyse firm survival and decision to internationalise (Carr, Haggard, Hmieleski & Zahra, 2010).

The discrete-time logit model has also been extensively used to model students outcomes and in particular, dropout (Alarcon & Edwards, 2013; Bruinsma & Jansen, 2009; Gury, 2011; Hovdhaugen, 2015; Lassibille & Navarro Gómez, 2008; Ishitani, 2006; Murtaugh et al., 1999; Willett & Singer, 1991). Estimates from the logit and probit models tend to be almost equivalent, while those from the complementary log-log model

can deviate substantially from those obtained from the probit and logit models (Box-Steffensmeier & Jones, 2004). This is often the case in data sets where there are relatively few failures. In general, there are no clear reasons to prefer one link function over the other. The discrete-time logit model is adopted for this study.

2.5.2 Competing risks models

Simple applications of survival analysis assume that individuals are at risk of only experiencing one event. However, in many applications, individuals are at risk of experiencing two or more events which affect each other. Single events models do not account for the possible interdependence between competing outcomes. The outcomes competing with the main event of interest are treated as censored and censoring is assumed to be non-informative (random). According to DesJardins (2003), estimating dropout as a single risk may result in model misspecification since dropout and graduation may be negatively correlated. Competing risks are related, but mutually exclusive dependent events. Effectively this means that experiencing one event excludes someone from the risk of experiencing another event during the period under observation (Allison, 1984). The observations/units of interest are exposed to different kinds of risks at the same time interval, but it is assumed that the eventual failure of an observation results from only one of these risks, which is called "a cause of failure".

Competing risks models were originally applied in the health, medical and actuarial sciences. Their applications have since broadened to other fields. For instance, Van Praag (2003) used a competing risks approach to investigate determinants of business survival and success. The study focused on two types of exits in business, namely: voluntary exit, attributed to a lack of motivation and willingness to continue in business; and compulsory exit, attributed to insufficient (financial) opportunity to continue in business. A compulsory exit was associated with business failure, whereas a voluntary exit was associated with business success. Gregory-Smith, Thompson & Wright (2009) also used a competing risks framework to model the tenure and type of exit of Chief Operating Officers (CEOs) from FTSE 350 companies during 1999-2005.

The interest was on the effect of several independent variables such as company performance as measured by total shareholder return, proportion of insiders on the board, number of directors appointed during the CEO's tenure, board size on the mode of exit (retirement, dismissal, or other exits). In the field of political science, de Rouen Jr & Sobek (2004) applied a competing risks approach to analyse duration of civil wars and factors associated with civil war outcomes of interest, i.e. government victory, rebel victory, truce or treaty. The study looked at 92 civil wars in 53 states that occurred between 1944 and 1997. Shaw (2011) used data from the Survey of Income

Program Participation to identify factors that influence transitions of individuals from cohabitation to marriage or being single. A competing risks survival analysis model was employed in the study. Diermeier & Stevenson (1999) used the model to estimate cabinet survival time and analyse determinants of government termination resulting in cabinet dissolution or cabinet replacement.

In general, when modelling competing risks, if the occurrence of a competing event is not of substantive interest in the study and the competing event may be assumed to be independent of the occurrence of the primary event of interest, then the observations associated with the competing non-primary events may be treated as censored. However, when this is not the case, the competing risk must be accounted for. According to Allison (2014), failure to distinguish competing events in survival analysis may produce spurious conclusions.

For example, DesJardins et al. (2002) investigated the impact of student demographic characteristics, attitudinal variables and financial aid variables on graduation at a University of Minnesota-Twin Cities campus for students registered for the first time in the fall semester of 1991. In the study, ability and academic performance were adjusted for, and graduation rate between different colleges was compared. A competing risk approach was used to estimate stopout after the first, second, third, fourth, fifth and sixth year and graduation in the fourth, fifth and sixth year after registration. The results indicate that when graduation is modelled alone, Latino students were less likely to graduate compared to white students. However, when graduation and stopout are modelled jointly, the Latino results are not as nearly strong as modelled independently. Accounting for the interrelationship between graduation and stopout thus reduces some of the negative relationship between graduation and being Latino. Cumulative grade point average (GPA) was also found to decrease graduation chances when modelled alone. In the competing risk model, GPA was found to be strongly and positively related to timely graduation.

Ortiz & Dehon (2013) investigated the extent to which socio-economic background and financial aid influence academic success using the competing risks survival analysis technique developed. The study focused on students enrolled at a Belgian university for the first time in the academic years 1997-1998 and 2001-2002. These students were followed over a twelve-year period. The results indicate that when graduation is treated as a competing risk, female students have a 25% lower hazard of dropping out versus staying enrolled. On the other hand, when dropout is treated as a competing risk, female students have approximately 57% higher chance of graduating from university. When it comes to nationality, the results show that foreign students are more likely to experience successive enrolments without getting a degree, whereas Belgian students

are enrolled, on average, for fewer periods than foreign students.

In a study conducted at the University of KwaZulu-Natal (UKZN) in South Africa, Murray (2014) used a competing risks approach to identify institutional and student specific factors influencing the type of outcome experienced by students when they leave the university. The study focused on the length of time it took students to either graduate or dropout of their studies. The study involved students enrolled for a degree at this institution between the years 2004 and 2012. One of the objectives of the study was to compare the time it takes to graduate with time to voluntary and involuntary (academic exclusion) dropout. The results of the study indicate that when involuntary and voluntary dropouts are treated as competing risks, students with some form of residence-based accommodation graduated more earlier than those without residence-based accommodation. Having some form of financial aid and higher matric point score also contributed to students to graduate more quickly. On the other hand, when graduation and voluntary dropout are treated as competing risks, the results indicate that having some form of financial aid and residence-based accommodation increased the length of time that students stayed in the system before dropping out involuntarily. Finally, when graduation and academic exclusion are treated as competing risks, access to some form of financial aid and residence-based accommodation assists in preventing students from dropping out voluntarily.

In a recent study, Zewotir et al. (2015) used a competing risk approach to determine factors that influence successful completion or dropout from a master's programme at the UKZN. Based on a seven-year period beginning in 2004, the study focused on factors associated with the actual number of years it took students to graduate, while treating dropout as a competing risk; and the number of years it took students to drop out, while treating graduation as a competing risk. The results show that 50% of master's students had either graduated or dropped out within two years of registration. In terms of financial aid, the results showed that receiving some form of financial funding appeared to reduce the length of time it takes a student to dropout from a master's programme. Moreover, receiving some form of financial aid also reduces the length of time it takes to successfully complete a master's programme. When looking at race, the results indicate that race has no significant effect on dropout. However, for the students who eventually graduated, the results show that time to graduation was shorter for African students.

In the current study, at risk students are followed until the occurrence of either a graduation or a dropout. Students who graduate are no longer at risk of dropout and students who dropout are also assumed to be no longer at risk of graduation since dropout is assumed to be permanent as per the definition adopted in the study.

2.5.3 Analysis of discrete-time competing risks models

A common approach for analysis of discrete-time competing risks is to model the cause-specific discrete hazard function. The cause-specific hazard function is the hazard function for each event type. The discrete-time competing risks cause-specific hazard is the instantaneous risk of experiencing an event of interest given that no other competing event has occurred. In the discrete-time case, the cause-specific discrete hazard function can be modelled through the use of regression models for multi-categorical response variable (Tutz, 1995). The cause-specific hazard can either be modelled simultaneously using a multinomial logistic model or separately treating all other competing events as censored. The two alternatives are discussed below.

When discrete-time competing events are modelled separately, simple survival analysis modelling approaches such as the Gompertz or Weibull for parametric survival models, or a semi-parametric model such as Cox PH model can be employed. Time to each event of interest is separately analysed, treating all the other events as censored. Under this approach, it is assumed that the competing events are independent conditional on a set of observed explanatory variables. This process is repeated by replacing the event of interest with other competing events, resulting in K different models, where K is the number of competing events. This in essence, is equivalent to the estimation of single event survival analysis models with random censoring. Since separate modelling of cause-specific discrete hazard assumes independence of event times, it is important that this assumption is tested. However, this assumption is impossible to test since the eventual event times of competing events that did not occur first cannot be observed (Klein & Moeschberger, 2006). Estimates of the cause-specific hazard rates and effects of explanatory variables on those hazards may be unreliable when the assumption of independence is enforced, while the underlying risks are indeed dependent. The results obtained under such conditions must therefore, be interpreted with caution to avoid misleading conclusions.

The multinomial logit (MNL) model is the most popular method for analysis of categorical responses. Consequently, the multinomial logistic regression is the most widely used method for modelling discrete-time competing risks hazards simultaneously (Tutz & Schmid, 2016; Scott & Kennedy, 2005). Examples include: Barnett et al. (2009) who investigated length of stay in hospital; Gibbons et al. (2003) who analysed waiting time to organ transplant; Moors & Bernhardt (2009) who investigated variables that influence the inclination of cohabiting couples to change their union from cohabitation to either marriage, separation or continued cohabitation; de Rouen Jr & Sobek (2004) who analysed duration of civil wars and factors associated with civil war outcomes of interest, i.e. government victory, rebel victory, truce or treaty; Kimber et al. (2010) who

examined survival and long term cessation of injection in a cohort of drug users; and Pennington-Cross (2010) in finance. In terms of student outcomes, Scott & Kennedy (2005) laid the foundation for its use in analysing dropout and degree attainment. Ortiz & Dehon (2013) used this method to investigate factors that influence both dropout and degree completion in lower university levels at a university in Belgium. The approach was also used by Van Der Haert et al. (2014) to identify determinants of time to dropout from doctoral studies and time to PhD completion, and many others.

MNL is an extension of the binary logistic model that allows the modelling of effects of covariates with more than two nominal outcomes. The multiple discrete cause-specific hazards are related to covariates as an MNL model. In a multinomial competing risks model, censored observations can be treated as a reference category, and the risks of each type of outcome of interest relative to the risk of not experiencing an outcome of interest is estimated simultaneously. For instance, in the case of K competing events, the MNL model estimates K-1 logit models to obtain parameter estimates on the type specific or destination specific hazards. Under this approach, the depended variable is treated as a polytomous qualitative choice variable. The MNL model avoids the proportionality assumption invoked in Cox PH model. It also allows for direct competition among the competing events. Estimation of parameters is done through the maximum likelihood method and the parameters are interpretable as logit coefficients. When all the explanatory variables are categorical, estimation of the MNL model can be easily performed through log-linear methods (Allison, 1982).

On the downside, the MNL model does not allow for correlations among competing risks. On the contrary, independence from irrelevant alternatives (IIA) is assumed. The IIA property allows researchers to estimate the parameters of the MNL model consistently using subset of alternatives, since elimination of irrelevant alternatives does not affect the odds of probability of the remaining alternatives. Moreover, unlike the likelihood function for the continuous-time model, the discrete-time likelihood cannot be factored into separate components for each of the competing events (Allison, 1982). This in turn means that the model for event 1 cannot be fitted separately from a model for event 2. The maximum likelihood estimation must hence be done simultaneously for all kinds of events (Allison, 1982), meaning that misspecification for event 1 may affect inferences about the model for event 2, and vice versa.

2.6 Unobserved heterogeneity

In some studies, there may be factors other than the measured covariates that may have a significant effect on the distribution of survival time. This is often referred

to as heterogeneity of the observation. In most studies, it is often not possible to account for all possible explanatory variables. Relevant covariates may be left out because they are unmeasurable, unobservable, or because the researcher may not be aware that they affect the outcome variable. It is well known that failure to control for unobserved heterogeneity in statistical modelling results in inconsistent inference (Van den Berg, 2001; Trussell & Richards, 1985). The effects of failure to control for unobserved heterogeneity have been extensively studied for continuous time survival data. It has been shown that failure to control for unobserved heterogeneity produces biased estimates in single risk models (Van den Berg, 2001; Heckman & Singer, 1984a; Lancaster, 1985, 1979) as well as competing risks survival analysis models (Van den Berg, 2001; Butler et al., 1989).

Vaupel, Manton & Stallard (1979), showed that uncontrolled heterogeneity results in positively biased longevity prediction estimates and negatively biased differences in mortality rates between different populations. Furthermore, it has been shown that failure to control for unobserved heterogeneity resulted in negative duration dependence such that increasing event hazard probabilities were diminished over time and decreasing hazard probabilities were accelerated over time (Keiding et al., 1997; Hougaard, 1995; Trussell & Richards, 1985; Heckman & Singer, 1984a). Even though most studies focused on continuous-time survival data, similar results have been reported for discrete-time survival data (Zhang, 2003; Baker & Melino, 2000; Nicoletti & Rondinelli, 2010).

In the case of student dropout, DesJardins (2003) explained that the hazard rates observed in the presence of unobserved heterogeneity may be as a result of simple variation in the risk of dropout across individuals that is related to their varying, but unobserved characteristics. According to DesJardins (2003), this may happen when students with a higher likelihood of dropout do so early in their studies and are subsequently removed from the risks set. For example, since research suggests that poor social and/or academic integration, lack of motivation, being first-generation student (FGS) is associated with dropout, students who are likely to remain in the risk set (remain enrolled) may be more motivated or have strong support system (DesJardins, 2003). These unobserved differences may result in false inferences about the temporal risk of dropout if not properly adjusted for.

DesJardins (2003) considers the case where motivation is an unobserved factor that may cause differences in the observed hazard of dropout to illustrate how uncontrolled heterogeneity may introduce bias in the results. For the purpose of illustration, the hazard rate of dropout for the motivated and unmotivated students is assumed to be constant, although not the same. DesJardins (2003) shows that even though the

likelihood of dropout is constant for each of these two groups, the overall hazard profile for all students will tend to decline. The decline is attributed to the continuous change in the proportion of both groups in the risk set, despite the constant hazard rates for both groups, i.e. high risk (unmotivated) students experience dropout early, such that as time passes, the risk set disproportionately consists of students who have lower dropout rates (higher survival rates). Since the risks sets disproportionately consist of motivated students who have a higher survival rate, hence a lower hazard of dropping out, the overall hazard of dropout will tend to decline over time.

2.6.1 Frailty model

Frailty models or random effects models are the standard approach used to account for unobserved heterogeneity that occur because some observations are more fail prone and hence more frail than other observations in the data (Box-Steffensmeier & Jones, 2004; Flinn & Heckman, 1982). The term frailty was first introduced by Vaupel, Manton & Stallard (1979) to refer to individual differences in longevity. Frailty models are the survival analysis equivalent to regression models, which account for unobserved heterogeneity and random effects (Gutierrez, 2002). Single-risk duration models can, therefore, be extended to account for unobserved heterogeneity by introducing a random component or frailty term into the model. The standard approach is to assume a functional form for the distribution of the frailty term (Heckman & Singer, 1984b). All observations are assumed to have different frailties which results in a change in individual hazards so that all observed individuals in the study are subject in principle to different levels of risks.

The use of the class of positive stable distributions is recommended by Hougaard (2000), as they have more desirable properties regarding the marginal distribution of the survivor function. The most widely used are the Gamma and Gaussian distributions (Box-Steffensmeier & Jones, 2004). However, the Gamma is more popular with continuous-time models due to analytical convenience Lancaster (1979) and theoretical reasons Van den Berg (2001). In the case of discrete-time single risk models, the assumption of a Gaussian distribution may be computationally convenient Hess & Persson (2012). Based on this assumption, the hazard models can be estimated as binary choice models with normal random effects using commonly available software packages.

The choice of distribution for the unobserved heterogeneity term has been widely studied in survival analysis. There is evidence indicating that parametric maximum likelihood estimates are sensitive to the functional form of the assumed distribution of the unobserved heterogeneity (Hougaard, 2000; Keiding et al., 1997; Heckman & Singer, 1984a; Vaupel et al., 1979). The extent of the asymptomatic bias depends on the dis-

crepancy between the true distribution of frailty and the assumed distribution. Heckman & Singer (1984a) showed through theoretical and empirical examples that the parametric maximum likelihood estimates are inconsistent if the distribution of the unobserved heterogeneity is miss-specified.

For the discrete-time case, the findings have been mixed. For instance, Baker & Melino (2000) found that misspecification of the heterogeneity distribution can result in substantially biased estimated parameters. Nicoletti & Rondinelli (2010) show that choosing a Gaussian distribution for the frailty term when the true one is discrete or Gamma, does not affect estimated parameters. They suggest that the bias reported by Baker & Melino (2000) is as a result of ignoring the normalisation required, i.e. parameters cannot be taken by their surface value when comparing models such as the probit and logit as the response function is based on different means and/or variances. The findings of Nicoletti & Rondinelli (2010) are supported by Trussell & Richards (1985). The heterogeneity distribution is, therefore, assumed to be Gaussian in this study.

Unobserved factors may also introduce stochastic dependence among competing events. If the unobserved variable has an effect in determining the timing of several events, they will be correlated. Van den Berg (2001) explain that if the unobserved determinants are dependent across the risks, then the failure times are dependent given the regressors. There is enough reason to expect such dependence, specifically if the unit of observation is an individual whose behaviour may affect all hazard rates. Models that assume that competing risks are independent, i.e. censoring mechanism is non-informative, are generally by far the most common approaches to competing risks models (Gordon, 2002). This assumption is similar to IIA in competing risks. However, this assumption may be questionable and even unlikely. When the underlying risks are indeed dependent, but dependence is imposed, the estimated parameters may be inconsistent with artificially small standard errors (Gordon, 2002). Depended risks can, therefore, be modelled using frailty or random effects models. The frailty model is the conventional method to account for dependence among competing events (Gordon, 2002). Hence frailty models accomplish two tasks by modelling unobserved heterogeneity and accounting for the dependence structure for clustered or multiple duration times, respectively.

2.6.2 Cured model

One of the assumptions of the Cox's regression model is that all observations under study will eventually experience the event of interest, provided that the observation period is long enough(Sy & Taylor, 2000). This in essence means that if the observation period is long enough, then the probability of event occurrence will approach one(Tutz & Schmid, 2016). However, there are cases where some observations do not eventually

experience the event of interest. This is usually evidenced by a Kaplan-Meier (KM) curve that eventually levels off into a plateau instead of approaching zero (Price & Manatunga, 2001). Such observations are referred to as "cured" a term inherited from clinical trials. The terms split-population and long-term survivors are also used to describe such observations. Boag (1949), Berkson & Gage (1952) and Haybittle (1965) are some of the first researchers to address this topic. The associated modelling techniques have since come to be known as cured models, split-population models or mixture models. Cured models represent a particular form of heterogeneity in survival data. They can also be seen as a special case of frailty referred to as binary frailty, where the population is divided into a proportion who are at risk and those who are never at risk, with the never at risk group being referred to as long-term survivors (Wienke, 2010).

In this study, the possibility of having cured observations (students who do not drop out even if the study period is long enough) is highly unlikely due to student exclusion policies. Students who remain in the university beyond a certain period without graduating are likely to be excluded and subsequently classified as dropouts. On the other hand, if the event of interest was graduation, the possibility of having cured observations would be higher. In such instances the proportion of cured observations would be estimated and the survivor function of the observations that would eventually drop out would be adjusted accordingly, in the event that the proportion is non-ignorable.

2.7 Factors associated with student retention

Research investigating the reasons for student dropout indicates that there is seldom a single reason for student dropout. Students may have various reasons to drop out of their study programmes or continue with their studies. The picture is often complex with student dropout being attributed to a combination of inter-related factors. It is, therefore, beyond the scope of this section to provide a comprehensive review and a detailed discussion of all possible factors associated with academic outcomes in HE in South Africa. The discussion will mainly focus on variables that are available on the TUT database and some psychological factors even though they are not available (not measured) for the study. The discussion will further be narrowed to studies conducted in South Africa so as to focus on those factors that have been found to be significant in the South African context.

In the absence of a single measure used universally to measure student academic outcomes/success, different approaches have also been used to analyse students' academic success/outcomes in South Africa. Some studies looked at academic success as a con-

tinuous variable by using student marks as a measure of success (Bokaba & Tewari, 2014; Goodman et al., 2011; Sommer & Dumont, 2011), whereas other studies focused on perceptions of students and lectures with regards to academic success or failure in higher education (Steenkamp, Baard & Frick, 2009; Zulu, 2008; Fraser & Killen, 2005). Some other studies looked at academic outcomes from a binary point of view by focusing on failure/pass, graduation or dropout or stopout (Baard, Steenkamp, Frick & Kidd, 2010; Breier, 2010; Letseka, 2009; Lourens & Smit, 2003). Different student theories of attrition have also been tested (Wawrzynski, Heck & Remley, 2012; Strydom, Mentz & Kuh, 2010; Petersen, Louw & Dumont, 2009). Against this background, the discussion on factors associated with retention is not limited to studies that formulated student success as a binary outcome. Instead, all the studies that focus on students' academic outcomes from different angles in HE in South Africa are reviewed so as to leverage on the broad knowledge generated from these studies.

Previous studies have highlighted the role of gender as one of the determinants of students' academic outcomes. Jacobs (2015) found that gender as a single variable is a better predictor of first year university success than Grade 12 average marks of first-time entering undergraduate students, with a prediction value of 29.7%. Zewotir et al. (2011), on the other hand, found the gender effect on failure rates of first year students to only be significant for students in the Faculty of Education and Health Sciences. According to Rooney (2015), being female increases the likelihood of graduation.

Financial support or lack thereof has also been cited as one of the main contributors to student dropout (Moeketsi & Mgutshini, 2014; Pocock, 2012; Letseka, 2009). In fact, running out of funding has been ranked as the number one reason for dropping out of university by Black African student (Letseka & Breier, 2008). According to Murray (2014), having some form of financial aid increased the length of time to dropout. Rooney (2015) found that being ineligible for financial aid increased the likelihood of graduation.

The role played by psychological factors such as motivation (Goodman et al., 2011; Sommer & Dumont, 2011; Petersen et al., 2009), self-esteem (Seabi, 2011), adjustment (Petersen et al., 2009) and student engagement in student academic success, have also been highlighted. Language is also one of the most important factors raised in discussions on academic success in HE (CHE, 2010). For instance, Rooney (2015) found that being proficient in English increased the likelihood of graduation. Van Rooy & Coetzee-Van Rooy (2015) suggest that language measures like Matric language marks and scores and academic literacy tests used by some universities are good predictors of academic success at a university.

The relationship between academic success in higher education and Matric performance,

has also been extensively studied. According to Rooney (2015), good high school grades increase the likelihood of graduation at university. Breier (2010) found significant differences between graduates and dropouts when compared on the basis of their Matric results. Visser & Hanslo (2005); Lourens & Smit (2003); and Maree, Pretorius & Eiselen (2003) found Grade 12 results to have a significant effect on students' first year success. In terms of engineering, Maree et al. (2003) found that students who passed first year engineering differed significantly from those who failed on study attitude in mathematics, problem–solving behaviour in mathematics and calculations subtests scores. Jawitz (1995) found that the Grade 12 physical science mark has the strongest positive correlation with first year engineering students marks.

The prevalent differences in academic performance between students of different races, remains one of the challenges facing HE. Various studies suggest that race is a significant determinant of academic success at University (Rooney, 2015; Murray, 2014). The findings by Sampson (2011) indicate that there is a significant strong association between graduation rates and race, such that African students had the lowest graduation rate followed by the coloured students, Indians and white students. There is also evidence to suggest that student accommodation is also one of the key factors that have a significant effect on students' academic performance (Murray, 2014; Zewotir et al., 2011).

In the South African context, students who are the first in their family to attend higher education (first-generation students), are more likely to be black Africans (Blacks, coloured and Indians), consequently most of the challenges faced by this group of students overlap with those experienced by Black students in HE education. They are vulnerable to unique transitional and developmental challenges as a result of their disadvantaged socio-economic and educational background due to inequalities of apartheid (Heymann & Carolissen, 2011). First-generation students (FGS) are often socially located in ways which disadvantage or reify their inherited identities. Siyengo (2015) highlighted difficulties in accessing higher education and financial support as one of the negative experiences of FGS in HE. The student residence experience, diversity at lecture halls, institutional culture as well the language of teaching and learning, are some of the other challenges experienced by this group of students (Siyengo, 2015). It is, however, acknowledged that not all Black students are FGS, hence the need to look at this factor separately.

2.8 Chapter summary

This chapter began by reviewing the definition of student dropout. This was followed by a discussion on student dropout theories and previous student dropout studies, with an emphasis on the statistical analysis approaches used in the studies. Literature related to censoring was also reviewed. Survival analysis models were introduced in more concrete terms that the introductory discussion in Chapter 1. Progression of one method to another was outlined, indicating the strength and weaknesses of the techniques used to model student dropout. Special attention was paid to the discrete-time single risk and competing risk models as the proposed models for the study. A selection of literature related to frailty and cured models was also reviewed. The chapter concluded with a brief discussion of the factors associated with student academic outcomes in HE is South Africa. A selection of literature related to frailty and cured models was also reviewed. The chapter concluded with a brief discussion of the factors associated with student academic outcomes in HE is South Africa.

A selection of factors discussed that are available on the Integrated Tertiary System (ITS) are used as covariates in the single risk and competing risk discrete-time models proposed for the study in Chapter 3 provides a detailed outline of the theoretical background of the discrete-time single risk and competing risk models proposed for the study. A theoretical background of the frailty model that accounts for unobserved heterogeneity in the single risk case is also presented in Chapter 3.

Chapter 3

Methodology

3.1 Introduction

This chapter provides a theoretical background of the discrete-time survival analysis models used in the study. The chapter is subdivided as follows. Section 3.2 outlines the main functions used in survival analysis both for continuous-time and discrete-time survival data. The continuous-time functions are presented as a foundation for the discrete-time case. The discrete-time hazard model is presented in Section 3.3. Section 3.4 discusses the competing risks model. The model used to account for unobserved heterogeneity in the discrete-time single risk model is presented in Section 3.5. The chapter concludes with a discussion on model assessment and diagnostics methods in Section 3.6.

It is assumed that survival time is represented by a non-negative random variable T. The distribution of the random variable T is generally characterised by three elementary functions, namely the probability density function (continuous-time case) or probability mass function (discrete-time case), the survivor function and the hazard function. Any of the three functions can be uniquely determined if any of them is known as shown in Section 3.2.3. In this study, survival time is assumed to be discrete. It is also assumed that the distribution of the survival data is unknown and as such the focus is on nonparametric discrete-time survival analysis techniques.

We further assume that we have a sample of n independent observations ($i = 1, 2, \ldots, n$). An observation is observed beginning from some natural starting point t=0. The appropriate starting point is clear in most cases. For instance, if the event of interest is student dropout, the obvious starting point is the date of enrolment. Assuming that there is an observed starting point for each observation, then an observation continues until time t_i, at which point either an event occurs, or the observation is censored. It is assumed that censoring is independent of the occurrence of events, i.e. observations

or individuals are not selectively withdrawn from the sample because they are more or less likely to experience an event (Allison, 1982). The actual survival time of a unit is the realisation or value of T, which may be denoted as t.

3.2 Basic survival functions

3.2.1 Probability density function and cumulative distribution function

Continuous-time

Let T be a continuous random variable representing survival times. The possible values of T have a probability distribution that is characterised by probability density function (PDF), $f(t)$, and cumulative distribution function (CDF), $F(t)$ (Box-Steffensmeier & Jones, 2004). The cdf of the random variable T is given by

$$F(t) = \int_0^t f(x)dx = P\{T \leq t\}, \tag{3.2.1}$$

which specifies the probability that the survival time T is less than or equal to some value t (Box-Steffensmeier & Jones, 2004). On the other hand, the pdf for all points that $F(t)$ is differentiable is given by

$$f(t) = \frac{dF(t)}{d(t)} = F'(t).$$

This in turn implies that

$$f(t) = \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t}, \tag{3.2.2}$$

giving the unconditional failure rate of event occurrences in an infinitesimally small differentiable area. This can also be seen by expressing f(t) in terms of probability. By the definition of the CDF in 3.2.1, the probability density can be written as

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t)}{\Delta t}. \tag{3.2.3}$$

Equation 3.2.3 gives the instantaneous probability that an event will occur (or a unit will fail) in a negligible small area bounded by t and $t+\Delta t$. Equations 3.2.2 and 3.2.3 show that the pdf is an unconditional failure rate. Either $F(t)$ or $f(t)$ can be used to specify an equivalent distribution. If $F(t)$ is differentiable, then $f(t)$ must exist. Therefore, either the PDF or CDF can be used to characterise the distribution of failure times.

Discrete-time

The notation for discrete-time data suggested by Singer & Willett (1993) is used in this section as well as the mathematical formulations outlined for survival analysis functions for discrete-time data. In the discrete-time case, events are recorded in discrete intervals, that is, the continuous time is divided into an infinite set of contiguous time periods:

$$(0,t_1],(t_1,t_2],\ldots,(t_{j-1},t_j],[tj,\infty).$$

Let j index time periods such that the j^{th} period begins immediately after time t_{j-1} and ends and includes time t_j. Let T be a discrete random variable that indicates the time period j when the event of interest occurs for a randomly sampled observation from the population. Then $T = t_j$ can also be expressed as $T = j$, such that the interval $(0,t_1)$ refers to $T = 1$ and similarly $T = j$ refers to the interval $(t_{j-1},t_j]$, meaning that the event occurred in the j^{th} interval. If we further assume that we have $i = 1,2,\ldots,n$ observations, then $T = t_j$ indicates that an event occurred at time t_j, where $j = 1,2,\ldots$ and $0 < t_1 < t_2 < \ldots$. Implied in the above statement is that t_{j-1} indicates the duration of non-occurrence of the event.

For a discrete random variable T, the probability mass function is given by

$$f(t_j) = P\left(T = t_j\right). \tag{3.2.4}$$

3.2.2 Survivor function

The distribution of T can also be characterised through the survivor function, which is the complement of the cdf. It gives the probability that the event of interest has not occurred by duration t. Survival probabilities, therefore, represent the proportion of the original sample that has not experienced the event of interest by time t while the survivor function refers to the chronological pattern of these probabilities. The survivor function cumulates the period-by-period risks of event non-occurrence together to evaluate the probability that a selected observation will not experience the event of interest.

Continuous-time

According to Cleves (2008), for a continuous random variable T, the survivor function is given by

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t).$$

The survival function has the following properties:

- $S(0)$=1
- $\lim_{x\to\infty} S(t)$=0
- $S(t)$ is a monotonic, non-increasing and non-negative function
- Plotting $S(t)$ against t gives the survival curve.

The survival function may be estimated by the Kaplan-Meier (KM) estimator, also referred to as the Product Limit Estimator. The KM estimator was first introduced by Kaplan & Meier (1958). It incorporates data from both censored and uncensored observations. Each observation contributes information, provided that the observation has not experienced the event of interest. Observations that do experience the event of interest contribute to the risks set until they experience the event of interest. Similarly, censored observations contribute to the risk set until the end of the observation period or are lost to follow-up. The KM estimator of the survivor function is defined as follows:

$$\hat{S}(t) = \prod_{t_j \leq t} \left[1 - \frac{d_j}{n_j}\right],$$

where d_j is the number of events observed at time j and n_j is the number of observations at risk of experiencing the event at time j. The estimated variance of these estimates can be obtained through the Greenwood's formula as follows:

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}^2(t) \sum_{t_j \leq t} \left[\frac{d_j}{n_j\left(n_j - d_j\right)}\right].$$

Discrete-time

In terms of a formal definition, the survival probability for observation i in time period j denoted by $S(t_{ij})$ is the probability that an observation i will survive past time period j (Singer & Willett, 2003), i.e.

$$S(t_{ij}) = P(T_i > j) = 1 - P(T_i \leq j).$$

The set of $S(t_{ij})$ for an observation is the observation's survival function. When the observations are not distinguished on the basis of covariates, the subscript i can be dropped and survival function for a random observation in the population can be written as $S(t_j)$ or $S(t)$. The plot of $S(t)$ against t is a non-increasing step function which jumps downwards at $t_1, t_2, \ldots$. The event must not have happened at period j or any prior time period. The chronological pattern of survival probabilities expressed as a function of time gives the survivor function. The value of the survivor function at the beginning of time is 1 as no observation has yet experienced the event of interest. Over

time, as events occur, the survivor function declines towards 0. When the hazard is high, the survivor function declines rapidly and when the hazard is low the survivor function declines slowly. Unlike the hazard function which can increase, decrease or remain constant between adjacent intervals, the survivor function never increases. During time periods where no event occurs, the survivor function remains constant.

The sample survivor function can be estimated as

$$\hat{S}(t_j) = \hat{S}(t_{j-1})[1 - \hat{h}(t_j)]. \tag{3.2.5}$$

3.2.3 The hazard function

The most common representation of the event time distribution is the hazard function (also known as the hazard rate or intensity). The hazard function is the most important element in survival analysis as it shows the risk of event occurrence at each time period and it also gives an estimate of when an event is likely to occur. Its magnitude in each time period indicates the risk of the event occurrence in that period. The greater the hazard, the higher the risk. It gives the rate at which observations fail by t, given that the observation survived until time t.

Continuous-time

For continuous-time survival data, the hazard function is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{3.2.6}$$

Equation 3.2.6 is the rate of failure per time unit in the interval $[t, t + \Delta t]$, conditional on survival at or beyond time t (Cleves, 2008). The numerator in equation 3.2.6 is the conditional probability that an event will occur in the interval $[t, t + \Delta t)$ given that it has not occurred before, while the denominator is the width of the interval. Dividing one by the other gives a rate of event occurrence per unit of time. Taking the limit as the width of the interval approaches zero, we obtain an instantaneous rate of occurrence. An increase in the hazard rate implies that the likelihood of failure increases as time passes, whereas a decrease indicates that the likelihood of failure decreases as time passes.

The rate can also take other forms over time, such as increase and then decrease, or decrease and then increase. The relationship between the hazard rate and survival probability can be used to estimate the sample survivor function. Information about survival can be deduced from the hazards probabilities since for each interval, the estimated hazard probability provides information not only about event occurrence

but also about the probability of non-occurrence. The hazard rate, survivor function, and density and distribution functions are therefore mathematically linked. If any one of these is specified, the others can be fully determined. In order to demonstrate these relationships, we take note that 3.2.6 can be expressed as

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P\left[(t \leq T < t+\Delta t) \cap (T > t)\right]}{P(T > t)\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t)}{\Delta t} \times \frac{1}{P(T > t)} \\ &= f(t) \times \frac{1}{S(t)}. \end{aligned} \tag{3.2.7}$$

This implies that the hazard rate captures the relationship between failure times and the survival function in the following way:

$$h(t) = \frac{f(t)}{S(t)},$$

where

$$f(t) = \frac{-dS(t)}{d(t)}.$$

In other words, the rate of occurrence of the event at duration t equals the density of events at t, divided by the probability of surviving to that duration without experiencing the event. This means the hazard rate can be equivalently written as

$$h(t) = \frac{\frac{-dS(t)}{d(t)}}{S(t)},$$

which is equivalent to

$$h(t) = \frac{-d\log S(t)}{d(t)}. \tag{3.2.8}$$

By intergrating 3.2.8 using $S(0)$=1, the survival function can be written as

$$S(t) = e^{-(\int_0^t h(u)du)},$$

which can be expressed as

$$S(t) = e^{-(H(t))},$$

where the term

$$H(t) = \int_0^t h(u)du, \tag{3.2.9}$$

is called the cumulative hazard rate. From 3.2.3, $H(t)$ can be written in terms of the survival function:

$$H(t) = -\log(S(t)),$$

and the density function can be written in terms of the cummulative hazard rate:

$$f(t) = h(t)e^{-H(t)}.$$

Discrete-time

For a discrete random variable T, the hazard rate at time t_j is defined as the conditional probability that a randomly selected individual from the population will experience the event of interest in time period j, given that they have not experienced it in any earlier time period (Singer & Willett, 1993), i.e.

$$h(t_j) = P\left(T = t_j \mid T \geq t_j\right). \tag{3.2.10}$$

Equivalent to the continuous-time hazard, the discrete hazard can be expressed as

$$\begin{aligned} h(t_j) &= P\left(T = t_j \mid T \geq t_j\right) \\ &= \frac{P\left[\left(T = t_j\right) \cap \left(T \geq t_j\right)\right]}{P\left(T \geq t_j\right)} = \frac{P\left(T = t_j\right)}{P\left(T \geq t_j\right)} \\ &= \frac{P\left(T = t_j\right)}{P\left(T > t_{j-1}\right)} = P\left(T = t_j \mid T > t_{j-1}\right) \\ &= \frac{f(t_j)}{S(t_{j-1})}. \end{aligned} \tag{3.2.11}$$

We note that since $f(t_j) = S(t_{j-1}) - S(t_j)$, equation 3.2.11 can be expressed as

$$h(t_j) = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}.$$

This relationship is important in obtaining an estimate of the survival function in the presence of censoring. By rearranging the last equation, we get the following

$$S(t_j) = S(t_{j-1})[1 - h(t_j)]. \tag{3.2.12}$$

We note that

$$S(t) = \frac{S\left(t_1\right)}{S\left(t_0\right)} \times \frac{S\left(t_2\right)}{S\left(t_1\right)} \times \cdots \times \frac{S\left(t_j\right)}{S\left(t_{j-1}\right)}.$$

This implies that equation 3.2.12 can be expressed as

$$S(t) = \prod_{t_j \leq t} \frac{S\left(t_j\right)}{S\left(t_{j-1}\right)} = \prod_{t_j \leq t} \left[1 - h\left(t_j\right)\right].$$

The idea of conditionality is innate in the definition of the hazard. This means that an observation i can only experience the event of interest in time period j if, and only if they have not experience it in any earlier period. Conditionality ensures that hazards represent the probability of event occurrence for those observations that are eligible to experience the event in that period, i.e. those in the risk set. Observations that experience the event of interest are removed from the risk set and are ineligible to experience the event in later periods.

The conditionality also ensures that the hazard probability of observation i in time period j evaluates their unique risk of event occurrence in that period. Each observation has a hazard function that describes their true risk of event occurrence over time that is distinguished to that of other observations on the basis of covariates. We can drop the subscript i that indexes observations, since we are only describing the distribution of event occurrence for a random sample of observations from a homogeneous population among whom we are not yet distinguishing. If we let $nevents_j$ denote the number of observations that experience the event of interest in time period j and $natrisk_j$ denote the number of observations at risk in time period j, then the discrete-time hazard in time period j can be estimated as:

$$\hat{h_j} = \frac{nevents_j}{natrisk_j}.$$

The magnitude of the hazard in each time period, indicates the risk of event occurrence in that interval. The estimated discrete hazard function can be examined by plotting its values over time. Instead of plotting the discrete-time hazard functions as a series of lines joined together as a step function, the discrete-time hazard probabilities are plotted as a series of points joined together by line segments (Singer & Willett, 2003). The plots can be used to identify periods of high risk and to characterise the shape of the hazard function (Singer & Willett, 2003).

The KM estimator can also be used to estimate the cumulative hazard function. The Nelson-Aalen estimator is another estimator used to estimate cumulative hazard. Compared to the KM estimator, it has better low sample properties (Moeschberger & David, 1971). The Nelson-Aalen estimator is defined as follows:

$$\hat{H}\left(t\right) = \prod_{t_j \leq t} \left[1 - \frac{d_j}{Y_j}\right].$$

Median lifetime

It is often of interest to identify the centre of the distribution of event times. In the absence of censoring, all event times would be known and a sample mean can be computed. However, in the presence of censoring, the median lifetime can be used as a measure of centrality. The estimated median lifetime is the value of T for which the value of the estimated survivor function is 0.5. It is the time period by which 50% of the sample has experienced the event of interest and 50% has not.

3.3 Discrete-time hazard model

3.3.1 Model estimation

The conditional probabilities at time t_j expressed in equation 3.2.10 which can be denoted by h_j, are the essential parameters of the discrete-hazard model, and as such the goal of discrete-time survival analysis is to estimate these conditional probabilities and their dependence on selected covariates (Singer & Willett, 2003). The values of the conditional probabilities, h_j lie between 0 and 1 since they are probabilities. The main interest is in investigating whether the risk of an event's occurrence differs systematically across different types of individuals or observations with their specific covariates. Different heterogeneities from explanatory variables are therefore, considered in the hazard model.

In order to introduce observed heterogeneity into the definition of the discrete-time hazard, we let $X_p(p = 1,2,3,\ldots,P)$ be a set of P covariates, of which each characterises the members of the population on a specific dimension. Since the values of some covariates vary with time, the values of these covariates are recorded in each time period. We let $x_{ij} = [x_{1ij}, x_{2ij}, \ldots, x_{Pij}]$ represent a vector of observation i's values for each of the P covariates in time period j. The subscript i is introduced into the definition of the population discrete-time hazard in 3.2.10 so that h_{ij} represents the conditional probability that observation i, with associated covariates $x_{ij} = [x_{1ij}, x_{2ij}, \ldots, x_{pij}]$ experiences the event of interest in time period j, given that they have not experienced in earlier time periods, i.e.

$$h_{ij} = P(T_i = j \mid T_i \geq j, X_{1ij} = x_{1ij}, X_{2ij} = x_{2ij}, \ldots, X_{pij} = x_{pij}). \tag{3.3.1}$$

Equation 3.3.1 indicates that the hazard is a function of each observation's values on a vector of covariates. However, the functional form of the dependence is not specified. The probabilities, h_{ij} are parametised to have a logistic dependence on the covariates and time periods as proposed by Allison (1982) and Cox (1972).

The logistic distribution

Let $P(y_{ij}=1)=\lambda_i$ and $P(y_{ij}=0)=1-\lambda_i$ denote the probability of an event's occurrence and non-occurrence, respectively. We further assume that this probability is a function of covariates, x. According to Box-Steffensmeier & Jones (2004), the logit function has the following form

$$\log\left(\frac{\lambda_i}{1-\lambda_i}\right)=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}\cdots+\beta_p x_{pi}. \tag{3.3.2}$$

This function specifies λ_i in terms of the log-odds ratio of the probability of an event occurrence to the probability of a non-occurrence. The logit coefficients, β_k , are therefore interpreted in terms of their relationship to the log-odds of an event occurrence. The model can be expressed in terms of the odds of event occurrence using the exponential function, i.e.

$$\hat{\lambda_i}=\frac{e^{\beta' x}}{1+e^{\beta' x}}.$$

This gives the predicted probability of an event occurrence, i.e. $\hat{\lambda_i}$, where $\exp(\beta' x)$ represents the exponentiated logit parameters for a given covariate profile. The hazard rate is used as the independent variable in survival analysis, since it gives the risk of event occurrence at each time period given that it has not occurred at earlier time periods.

The logistic parametisation of the sample hazard function, h_{ij} is therefore, given by the following:

$$h_{ij}=\frac{1}{1+\exp^{-}\left((\alpha_1 D_{1ij}+\alpha_2 D_{2ij}+\cdots+\alpha_J D_{Jij})+(\beta_1 x_{1ij}+\beta_2 x_{2ij}+\cdots+\beta_P x_{Pij})\right)}. \tag{3.3.3}$$

Equation 3.3.3 can be expressed as

$$\frac{h_{ij}}{1-h_{ij}}=\exp^{(\alpha_1 D_{1ij}+\alpha_2 D_{2ij}+\cdots+\alpha_J D_{Jij})}\times\exp^{(\beta_1 x_{1ij}+\beta_2 x_{2ij}+\cdots+\beta_p x_{pij})},$$

where J is the length of the observation period, i.e. the last time period observed for any observation in the sample and $[D_{1ij},D_{2ij},\ldots,D_{Jij}]$ are a series of dummy variables with values $[d_{1ij},d_{2ij},\ldots,d_{Jij}]$ indexing time periods, $[\alpha_1,\alpha_2,\ldots,\alpha_J]$ are the intercept parameters and $[\beta_1,\beta_2,\ldots,\beta_p]$ are the slope parameters which describe the effects of the covariates on the baseline mode (Singer & Willett, 1993). Let j_i represent the last period when observation i was observed (at which point they either experienced the event of interest or was censored). The time-period dummies are defined the same way for all the observations. For instance, $d_{1ij}=1$ when $j=1$, and $d_{1ij}=0$ when j takes on any other value (2 through J); $d_{2ij}=1$ when $j=2$ and $d_{2ij}=0$ otherwise; and so on.

By taking the log transformation on both sides of the equation we get:

$$\log\left(\frac{h_{ij}}{1-h_{ij}}\right) = (\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij}) + (\beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_P x_{Pij}). \quad (3.3.4)$$

Equation 3.3.4 suggests that the conditional log-odds that the event of interest will occur in each time period j given that is has not occurred in earlier time periods is a linear function of constant term α_j specific to period j, and of the values of the covariates at period j multiplied by the appropriate slope parameters. Assuming that we have a random sample of $n(i = 1,2,\ldots,n)$, the discrete-time hazard model in equation 3.3.4 can be fitted and its parameters can be estimated. Estimators for the parameters $[\alpha_1,\alpha_2,\ldots,\alpha_J]$ and $[\beta_1,\beta_2,\ldots,\beta_p]$ and therefore h_{ij} can be estimated through maximum likelihood estimation. The procedure is outlined in the next section.

3.3.2 Constructing the likelihood function

To construct the likelihood function we follow the approach outlined by Singer & Willett (1993). We first assume that each observation in the sample is observed through each successive discrete-time periods until it experiences the event of interest or it is censored. Let

$$y_{ij} = \begin{cases} 0 & \text{if observation } i \text{ does not experience the event in period } j \\ 1 & \text{if observation } i \text{ experience the event of interest in period } j. \end{cases} \quad (3.3.5)$$

Since we are interested in the occurrence of a single non-repeatable event, the sequence of y values can only show one of two patterns. Either y_{ij} assume the value of zero in every time period that is observed during data collection, including the last one, indicating that that observation did not experience the event of interest during the observation period and was subsequently censored or it assumes the value of one for the specific time period that is it experienced the event of interest, and observation is terminated afterwards. Since the event is not repeated, for an uncensored observation, fewer y values may be required to describe the event history, they will consist of a series of zeros terminating in the value one. In addition to describing event occurrence by a sequence of y values, we also define a censoring indicator δ_i, where

$$\delta_i = \begin{cases} 1 & \text{if observation } i \text{ is censored} \\ 0 & \text{if observation } i \text{ is not censored.} \end{cases} \quad (3.3.6)$$

Since the hazard function is conditional, observations only contribute data at time period j if they experience the event at that time period or if they have not yet experienced the event of interest by that time period. If the event of interest does not occur, then the observation is right-censored and contributes to the data set a vector of zeros. The number of time periods, can therefore, vary across observations $(j = 1,2,\ldots,j_i)$ where j_i is the terminal time period for observation i, that is time period with the last non-missing value for observation i. The subscript i indicates that the terminal period can vary for each observation. If we assume that

censoring is random, then the likelihood function for the sampled data consists of two parts, that is the part for uncensored observations (the probability that an observation experienced the event of interest in time period j_i) and another part for censored observations. The expressions for each part are derived separately as suggested by Allison (1982).

We first look at the probability that an uncensored observation will experience the target event in time period j_i. This can be written as a product of terms, one per period describing the conditional probabilities that the event did not occur in periods 1 through $j_i - 1$, but occurred in period j_i, i.e.

$$\begin{aligned} P(T_i = j_i) = P(T_i = j_i \mid T_i \geq j_i) \times P(T_i \neq j_i - 1 \mid T_i \geq j_i - 1) \times \ldots \\ P(T_i \neq 2 \mid T \geq 2) \times P(T_i \neq 1 \mid T_i \geq 1), \end{aligned} \tag{3.3.7}$$

where the subscripts 1 and 2 index the first and second time periods. We know that

$$h_j = P(T = j \mid T \geq j) \tag{3.3.8}$$

and

$$P(T > j \mid T \geq j) = 1 - h_j. \tag{3.3.9}$$

Expressing 3.3.7 in terms of the hazard probability and in terms of the conditional survival probability in 3.3.9, the probability of failure, $f(t)$, can be written as

$$\begin{aligned} P(T_i = j_i) &= h_{ij_i} \times (1 - h_{i(j_i-1)}) \times (1 - h_{i(j_i-2)}) \cdots \times (1 - h_{i2}) \times (1 - h_{i1}) \\ &= h_{ij_i} \prod_{j=1}^{j_i-1} (1 - h_{ij}). \end{aligned} \tag{3.3.10}$$

This shows that the probability mass function is equal to the hazard probability multiplied by the product of the conditional survivor functions. From equation 3.2.7, $f(t) = S(t)h(t)$. This means that the survivor function must be equal to

$$\begin{aligned} Pr\{T_i > j_i\} &= (1 - h_{ij_i}) \times (1 - h_{i(j_i-1)}) \times (1 - h_{i(j_i-2)} \times \cdots \times (1 - h_{i2}) \times (1 - h_{i1}) \\ &= \prod_{j=1}^{j_i} (1 - h_{ij}). \end{aligned} \tag{3.3.11}$$

If we assume that the sampled observations are independent (given their $x_{1ij}, x_{2ij}, \ldots, x_{Pij}$ values), then the likelihood function is simply the product of the probabilities of observing the sample data, $P(T_i = j_i)$ in the case of uncensored observations ($\delta_i = 0$) and $P(T_i > j_i)$ in the case of censored observations ($\delta_i = 1$), i.e.

$$\mathcal{L} = \prod_{i}^{n} [P(T_i = j_i)]^{1-\delta_i} [P(T_i > j_i)]^{\delta_i}. \tag{3.3.12}$$

Substituting equations 3.3.10 and 3.3.11 into 3.3.12 yields the following

$$\mathcal{L} = \prod_{i}^{n} \left[h_{ij_i} \prod_{i=1}^{j_i-1} (1-h_{ij}) \right]^{1-\delta_i} \left[\prod_{j=1}^{j_i} (1-h_{ij} \right]^{\delta_i}.$$

Taking the logarithms of equation 3.3.2 gives us the following log-likelihood function

$$\log \mathcal{L} = \sum_{i}^{n} (1-\delta_i) \log \left[\frac{h_i j_i}{(1-h_{ij_i})} \right] + \sum_{i=1}^{n} \sum_{j=1}^{j_i} \log (1-h_{ij}). \tag{3.3.13}$$

Using the indicator variable y_{ij} defined above, 3.3.13 can be written as

$$\log \mathcal{L} = \sum_{i}^{n} \left[\sum_{j=1}^{j_i} y_{ij} \log \left(\frac{h_{ij}}{(1-h_{ij})} \right) + \sum_{j=1}^{j_i} \log (1-h_{ij}) \right].$$

This simplifies to

$$\log \mathcal{L} = \sum_{i}^{n} \sum_{j=1}^{j_i} \left[\log \left(\frac{h_{ij}}{(1-h_{ij})} \right)^{y_{ij}} + \log (1-h_{ij}) \right]. \tag{3.3.14}$$

Taking the antilog of equation 3.3.14 and combining like terms gives us

$$\mathcal{L} = \prod_{i}^{n} \prod_{j=1}^{j_i} h_{ij}^{y_{ij}} (1-h_{ij})^{1-y_{ij}}. \tag{3.3.15}$$

Equation 3.3.15 is the likelihood function for the discrete-time hazard. According to Singer & Willett (1993) and Allison (1982), this likelihood function is equivalent to the likelihood function for or a sequence of U ($U = j_1 + j_2 + \cdots + j_n$) independent *Bernoulli* trials with parameters h_{ij}. The parameters h_{ij} are the probability of the binary observed variables, y_{ij}. The discrete-time hazard model can therefore be estimated using binary regression models. The logit link function is used in the study in line with Singer & Willett (1993) to specify how the hazard rates depend on the explanatory variables.

This equation can be expressed as

$$\frac{h_{ij}}{1-h_{ij}} = \exp^{(\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij})} \times \exp^{(\beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij})}. \tag{3.3.16}$$

Using the likelihood function derived in equation 3.3.15, we have the following

$$\log \mathcal{L}(\theta) = \sum_{i}^{n} \sum_{j=1}^{j_i} y_{ij} \log \left[\frac{h_{ij}}{(1-h_{ij})} \right] + \log (1-h_{ij}), \tag{3.3.17}$$

where

$$\theta' = [\alpha_1, \alpha_2, \ldots, \alpha_J, \beta_1, \beta_2, \ldots, \beta_p]$$

and

$$h_{ij} = \frac{1}{\exp^{-\left(\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij}\right)}}.$$

Creating the person period data set

The steps to create a person period data set are outlined by Singer & Willett (2003). In the normal person-oriented data set, each observation in the sample has one record of the data. A record for the ith observation usually contains information about:

- Duration: the length of time an observation was observed, usually recorded as the last time period, j_i,in which an observation was observed.
- Censoring: the value of the censoring indicator δ_i, which indicates whether an observation experienced the target event in the last time period in which it was observed or whether the observation was censored. The censoring indicator δ_i has a value of 0 if observation i was not censored in time j_i and 1 if it was.
- Selected explanatory variables: observation i's values on covariates recorded in each time period j up to, and including, time period j_i. For time-invariant covariates, only a single value is recorded for all periods. However, time-varying covariates may take on a different value in each time period.

The person-oriented data set needs to be converted into a new person-period data set such that each observation has multiple records, one for each time period of observation. New variable musts be created to distinguish the multiple records within an observation. The new variables are referred to as time indicators. The values of the independent variables must be recorded such that they are appropriate to each period. An event indicator variable Y, is created using the duration and censoring information. In the new person-period data set, the ith observation has j_i records, with the jth of these containing information about the jth time period:

- The time indicators: the set of dummy variables $D_{1ij}, D_{2ij}, \ldots, D_{Jij}$ assume values that identify the particular time period to which the record refers to. All the time indicators take on the value 0, except for the jth dummy, D_{jij}, which takes the value 1.
- The independent variables: for the jth record, the independent variables contain the ith observation's values of the P covariates for the time period j, $X_{1ij}, X_{2ij}, \ldots, X_{Pij}$.
- The event indicator: the variable Y records values y_{ij} that indicate whether or not the target event has occurred for observation i at time period j. it takes on the value 1 if the event of interest has occurred and 0 otherwise.

The original n records of the person-oriented data set become $U = (j_1 + j_2 + \cdots + j_n)$ in the new person-period data set.

3.3.3 Inclusion of time-varying covariates

The discrete-time survival analysis model adapts naturally to the inclusion of time-varying predictors (Singer & Willett, 2003). Since the models are fit using a person-period data set, a time-varying predictor simply takes on its applicable value for each time-varying covariate. The model in 3.3.16 allows for inclusion of time-varying predictors since each variable has two subscripts: i indexing the observations and j indexing time periods. For example if we assume that we have two covariates, X_{1i} which is constant over time and X_{2ij}, the discrete-time model with a logit link can be expressed as

$$\text{logit}\,(h_{ij}) = [\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij}] + [\beta_1 X_{1i} + \beta_2 X_{2ij}]\,.$$

This means that an observation i's value of the logit hazard in time j depends on their value of X_1 which is constant across all time points and their value of X_2 at time period j.

3.4 Discrete-time competing risk model

3.4.1 Introduction

Let time assume values $\{1,\ldots,k\}$ we also let $q = k-1$. If it results from intervals, one has k underlying intervals $[a_0,a_1)$, $[a_1,a_2),\ldots,$ $[a_{q-1},a_q)$ $[a_q,\infty)$, such that it is typically assumed that $a_0 = 0$ and a_q denotes the final follow-up (Tutz & Schmid, 2016). Discrete time $T \in \{1,\ldots,k\}$ means that $T = t$ is observed if failure occurs within the interval $[a_{t-1},a_t)$. If it is inherently discrete, then T is the original observation (Tutz & Schmid, 2016).

3.4.2 Cause-specific discrete hazard function

The single risk discrete-time model represented by equation 3.3.3 can be extended to a MNL model, which is an extension of the logit model (Tutz & Schmid, 2016). It is essentially a series of linked logit model. The MNL model estimates $m-1$ logit models to obtain parametric estimates on the cause-specific or destination specific hazards (Tutz & Schmid, 2016).
Let the distinct terminating events be denoted by $R \in \{1,\ldots,m\}$. Then the cause-specific discrete hazard function resulting from risk or cause r is given by the conditional probability

$$\lambda_r(t \mid x) = P(T = t, R = r \mid T \geq t, x), \tag{3.4.1}$$

where x is a vector of covariates and $r = 1,2,\ldots,k$ and $t = 1,2,\ldots,q$. Time-varying and lagged covariates can be easily included in 3.4.1, i.e.

$$\lambda_r(t \mid x_t) = P(T = t, R = r \mid T \geq t, x_t), t = 1,2,\ldots, \tag{3.4.2}$$

where x_t may comprise time-constant or possibly lagged time-dependent covariates observed at time t.

The overall hazard function is obtained by summing the m $(\lambda_1(t \mid x), \ldots, \lambda_m(t \mid x))$ hazard functions, i.e.

$$\lambda(t \mid x) = \sum_{r=1}^{m} \lambda_r(t \mid x) = P(T = t \mid T \geq t, x). \tag{3.4.3}$$

3.4.3 Survival function

The survival function of an event and the unconditional probability of an event in period t_i have the same form as in the single risk case and are given by:

$$\begin{aligned} S(t \mid x) &= P(T > t \mid x) \\ &= \prod_{j=1}^{t} (1 - \lambda(j \mid x)) \end{aligned} \tag{3.4.4}$$

and

$$\begin{aligned} P(t \mid x) &= \lambda(t \mid x) \prod_{j=1}^{t} (1 - \lambda(j \mid x)) \\ &= \lambda(t \mid x) S(t-1 \mid x). \end{aligned} \tag{3.4.5}$$

If an individual survives until interval $[a_{t-1}, a_t)$ then there are $m+1$ possible outcomes, that is one of the m target events or survival beyond $[a_{t-1}, a_t)$. The corresponding resultant conditional probabilities are given by

$$(\lambda_1(t \mid x), \ldots, \lambda_m(t \mid x)), 1 - \lambda(t \mid x),$$

where $1 - \lambda(t \mid x)$ is the survival probability.
The multi-category models that are used in the modelling of categorical data are the natural models for the $m+1$ events. Therefore if an observation reaches interval $[a_{t-1}, a_t)$, a natural parametric model for the hazards in the MNL model given by

$$\lambda_r(t \mid x) = \frac{exp\left(\beta_{0tr} + x^T \gamma_r\right)}{1 + \sum_{s=1}^{m} exp\left(\beta_{0ts} + x^T \gamma_s\right)}, \tag{3.4.6}$$

where $t = 1, \ldots, q$ and $r = 1, \ldots, m$. The parameters $\beta_{01r}, \ldots, \beta_{0qr}$ are the cause-specific baseline hazard functions and γ_r represent the cause specific vector of coefficients. In a multinomial logit model, conditional survival corresponds to the reference category. It is, therefore, sufficient to specify the conditional probability of the target events $1, \ldots, m$. The conditional probability of survival is given by

$$\begin{aligned} P(T > t \mid T \geq t, x) &= 1 - \sum_{r=1}^{m} \lambda(t \mid x) \\ &= \frac{1}{1 + \sum_{s=1}^{m} exp(\ \beta_{0ts} + x^T \gamma_s)}. \end{aligned} \tag{3.4.7}$$

With $R \in \{1, \ldots, m\}$, where $R = 0$ denotes the conditional survival, the conditional probabilities are given by $\lambda_0(t \mid x) = P(T > t \mid T \geq t, x), \lambda_1(t \mid x) \ldots, \lambda_m(t \mid x)$, which add up to one.

3.4.4 Model estimation

The steps outlined by Möst, Pößnecker & Tutz (2016) are followed for model estimation. Let the data be given by $(t_i, r_i, \delta_i, x_i)$, where $i = 1, \ldots, n$, x_i is a vector of explanatory variables, $t_i = min(T_i, C_i)$ is the observed discrete time, which is the minimum of survival time T_i and censoring time C_i (Möst et al., 2016). It assumed that censoring is random, i.e. T_i and C_i are independent . We further assume that $r_i \in \{1 \ldots \ m\}$ indicates the target event, x_i is a covariate vector and δ_i is a censoring indicator such that

$$\delta_i = \begin{cases} 1 & T_i \leq C_i, \\ 0 & T_i > C_i. \end{cases}$$

From the above equation $r_i = 0$ if and only if $\delta_i = 0$. Implied in this definition of the censoring indicator is that censoring occurs at the end of the interval. The likelihood contribution of the i^th observation for the model 3.4.6 is given by

$$\mathcal{L}_i = P(T_i = t_i, R_i = r_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i} \tag{3.4.8}$$

It assumed that censoring is random,i.e. T_i and C_i are independent. We further define an indicator function δ_i such that
The conditioning on the vector of covariates x_i is omitted for notational simplicity. If we assume that censoring is random, then the censoring mechanism does not depend on the parameters that determine the survival time (Kalbfleisch & Prentice, 2002), the factor $c_i = P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ can be omitted in the above equation, resulting in the following

$$\begin{aligned} \mathcal{L}_i &= P(T_i = t_i, R_i = r_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} \\ &= \lambda_{r_i}(t_i \mid x_i)^{\delta_i} (1 - \lambda(t_i \mid x_i))^{1-\delta_i} \prod_{t=1}^{t_i - 1} (1 - \lambda(t \mid x_i)). \end{aligned} \tag{3.4.9}$$

We let $R_t = \{i : t \leq t_i\}$ be the risk set consisting of all observations/individuals who are at risk in interval $[a_{t_i - 1}, a_{t_i})$. For an alternative representation of the likelihood function, we define the following indicators for the transition to the next period

$$Y_{itr} = \begin{cases} 1, & \text{event of type } r \text{ occurs in the interval } [a_{t_i-1}, a_{t_i}) \\ 0, & \text{no event of type } r \text{ occurs in the interval } [a_{t_i-1}, a_{t_i}), \end{cases} \tag{3.4.10}$$

and

$$Y_{it0} = \begin{cases} 0, & \text{event of type } r \text{ occurs in the interval } [a_{t_i-1}, a_{t_i}) \\ 1, & \text{no event of type } r \text{ occurs in the interval } [a_{t_i-1}, a_{t_i}), \end{cases} \tag{3.4.11}$$

where $j \in R_t$ and $r = 1, \ldots, m$. The indicator variable in 3.4.11 can therefore be derived from 3.4.10 by $y_{it0} = 1 - y_{it1} - \cdots - y_{itm}$. The indicator variables can be gathered in the vector $y_{it}^T = (y_{it0}, y_{it1}, \ldots, y_{itm}) = (1, 0, \ldots, 0)$, which denotes the response vector of observations i, $i = 1, \ldots, n$ and $t = 1, \ldots, t_i$. The likelihood function for the i^{th} observation can be written as

$$\begin{aligned} \mathcal{L}_i &= \prod_{t=1}^{t_i} \left(\prod_{r=1}^{m} \lambda_r (t \mid x_i)^{y_{itr}} \right) (1 - \lambda (t \mid x_i))^{y_{it0}} \\ &= \prod_{t=1}^{t_i} \left(\prod_{r=1}^{m} \lambda_r (t \mid x_i)^{y_{itr}} \right) \left(1 - \sum_{r=1}^{m} \lambda_r (t \mid x_i) \right)^{y_{it0}}. \end{aligned} \tag{3.4.12}$$

This means that the likelihood for the i^{th} observation is the same as the likelihood for the t_i obervations $y_{i1}, \ldots, y_{it_i}$ of a multinomial response model. Given that an observation survives until interval $[a_{t-1}, a_t)$, the response is multinomially distributed with $y_{jt}^T = (y_{jt0}, y_{jt1}, \ldots, y_{jtm}) \sim M(1, 1 - \lambda(t \mid x_i), \lambda_1(t \mid x_i) \ldots, \lambda_m(t \mid x_i))$. The dummy variable $y_{ito} = 1 - y_{it1} - \cdots - y_{itm}$ has the value 1 if observation i does not experience the event of interest in interval $[a_{t-1}, a_t)$ and $y_{it0} = 0$ if observation i experiences the event of interest in interval $[a_{t-1}, a_t)$. Consequently, the likelihood is that of the multicategorical model

$$P(Y_{it} = r \mid x_i) = P(y_{itr} = 1 \mid x_i) = \frac{\exp(\eta_{itr})}{1 + \sum_{s=1}^{m} \exp(\eta_{its})},$$

where $\eta_{itr} = \beta_{0tr} + x_j{}^T \gamma_r$. The total log-likelihood is given by

$$\begin{aligned} l &= \sum_{i=1}^{n} \sum_{t=1}^{t_i} \left[\sum_{r=1}^{m} y_{itr} \log \lambda_r (t \mid x_i) + y_{it0} \log \left(1 - \sum_{r=1}^{m} \lambda_r (t \mid x_i) \right) \right] \\ &= \sum_{t=1}^{q} \sum_{i \in R_t} \left[\sum_{r=1}^{m} y_{itr} \log \lambda_r (t \mid x_i) + y_{it0} \log \left(1 - \sum_{r=1}^{m} \lambda_r (t \mid x_i) \right) \right]. \end{aligned}$$

As in the single risk case, the ML estimates may be computed within the framework of multivariate GLMs after construction of an appropriate design matrix. Let $\mathbb{1}_t = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ be a vector of length q with 1 in t^{th} position and zeros otherwise and let $\tilde{x}_{it}^T = (\mathbb{1}_t^T, x_i^T)$ denote a design vector that includes the baseline effect for time period t and the covariate vector x_i. With corresponding parameter vectors $\tilde{\gamma}_r^T = (\beta_{01r}, \ldots, \beta_{0qr}, \gamma_r^T) = (\beta_{0r}^T, \gamma_r^T)$, we can obtain for the linear predictors $\eta_{itr} = \beta_{0tr}^T + x_i^T \gamma_r^T$ the closed form

$$\eta_{it} = (\eta_{it1}, \ldots, \eta_{itm})^T = \left(\tilde{x}_{it}^T\tilde{\gamma}_1, \ldots, \tilde{x}_{it}^T\tilde{\gamma}_m\right)^T.$$

The matrix of linear predictors for all artificial data points that belong to one real observation is given in compact form by the following

$$\eta_j = \begin{bmatrix} \eta_{i1}^T \\ \vdots \\ \eta_{it_i}^T \end{bmatrix}_{t_i \times m} = \tilde{X}_i\tilde{\Gamma} = \begin{bmatrix} \tilde{x}_{i1}^T \\ \vdots \\ \tilde{x}_{it_i}^T \end{bmatrix}_{t_i \times (q+p)} \left[\tilde{\gamma}_1 | \ldots | \tilde{\gamma}_m\right]_{(q+p)\times m}.$$

Then for all the observations we have $\eta = \tilde{X}\tilde{\Gamma}$, with $\eta^T = (\eta_1^T | \ldots | \eta_n^T)$ and $\tilde{X}^T = (\tilde{X}_1^T | \ldots | \tilde{X}_n^T)$ so that estimation and inference is readily available via standard methods for multivariate GLM.

3.5 Unobserved heterogeneity

Estimation of the distribution of unobserved heterogeneity is done by assuming that the frailty term follows a Gaussian distribution with zero mean (Hess & Persson, 2012). It is assumed that that there exists different types of observations or individuals who differ between them in unobserved attributes, e.g. motivation, which affect dropout and graduation rates.

3.5.1 Single risk model

Unobserved heterogeneity can be accounted for by including random effects into the logit model. In the discrete time case, an individual specific unobserved random effects term u_{ij} for the j^{th} time period is introduced in the model to account for unobserved heterogeneity. From 3.5.1, the discrete time hazard model has a logistic dependence on explanatory variables and time, i.e.

$$h_{ij} = \frac{1}{1 + \exp^{-}(\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij})}.$$

Inclusion of the u_{ij} gives the following

$$h_{ij} = \frac{1}{1 + \exp^{-}(\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \cdots + \alpha_J D_{Jij} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + u_{ij})}. \quad (3.5.1)$$

3.6 Model diagnostics

The aim in statistical model developments is to obtain the model that best describes the central point of the data. In most cases, the results of the fitted model are summarised in terms of point and interval estimates of practically interpretable measures of the effect of explanatory variables on dependent variable. Examples of summary measures include, mean

differences for linear regression, the odds ratio for logistic regression and hazard ratio for proportional hazards regression. Inferences based on models depend entirely on the fitted statistical model. A thorough examination of the adequacy of the fitted statistical model is therefore crucial. It assures us that the inferences made based on the fitted model are the best and most valid possible. Goodness-of-fit tests are used to formally determine the adequacy of the fitted model.

3.6.1 Assessing overall goodness-of-fit

In survival analysis the interest is on the effect of explanatory variables on survival. The interest is in the contribution of predictors, i.e. whether the predictor variables in the model that account for the heterogeneity in the population are relevant. When maximum likelihood estimation is used to generate the log-likelihoods, then the log-likelihood (LL) is used to assess the fit of the model.

Likelihood ratio test

The likelihood ratio (LR) test is used when the interest is in comparing nested models. The interest is in testing hypothesis of the form:

$$H_o : C\theta = \xi \quad \text{against} \quad H_o : C\theta \neq \xi, \tag{3.6.1}$$

where C is a fixed matrix of full rank $s \leq p$ and ξ is a fixed vector. The vector θ in the linear hypothesis $C\theta = \xi$ collects all the parameters of the model, i.e. $\theta^T = (\gamma_{01}, \ldots, \gamma_{oq}, \gamma^T)$ (Cleves, 2008). A common test statistic for linear hypothesis is the likelihood ratio test, which is based on the comparison between two models, that is, the model without constraints and the model fitted under the linear constraints. Suppose that $\hat{\theta}$ denotes the maximum likelihood estimate for the full model in 3.3.16 and $\tilde{\theta}$ denotes the maximum likelihood estimate under the constraint $C\theta = \xi$. Then the likelihood ratio statistic is given by:

$$LR = -2\left[l\left(\tilde{\theta}\right) - l\left(\hat{\theta}\right)\right], \tag{3.6.2}$$

which measures the change of the log-likelihood given in 3.3.17 when evaluated as $\hat{\theta}$ and $\tilde{\theta}$, where $l\left(\tilde{\theta}\right)$ is the sample LL statistic of the current model and $l\left(\hat{\theta}\right)$ is the LL of the saturated model (Cleves, 2008). Under the regularity conditions LR statistic follows asymptotically a χ^2 with $s = rk(C)$ degrees of freedom, such that $s = rk(C)$ denotes the rank of matrix C.

The Wald statistics can be used as an alternative test statistic. It is derived as an approximation of the LR and it is given by:

$$w = \left(C\hat{\theta} - \xi\right)^T \left[CF^{-1}\left(\hat{\theta}\right)C^T\right]^{-1}\left(C\hat{\theta} - \xi\right), \tag{3.6.3}$$

where $F(\theta) = E(-\partial^2 l(\theta)/\partial\theta\partial\theta^T)$ denotes the Fisher information matrix. One advantage of the Wald statistic is that only the full model has to be fitted to obtain θ (Singer & Willett,

2003). It is not necessary to fit the constrained model. Another alternative test statistic is the score statistic which is given by :

$$u = \; s^T\left(\tilde{\theta}\right) F^{-1}\left(\tilde{\theta}\right) s\left(\tilde{\theta}\right), \tag{3.6.4}$$

where $s(\theta) = \partial l(\theta)/\partial\theta = (\partial l(\theta)/\partial\beta_1, \ldots, \partial l(\theta)/\partial\beta_p)^T$ is the score function evaluated at the fit of the constrained model.

The deviance statistic

The deviance statistic compares LL for two models: (1) the current model, which is the model just fit; and (2) a saturated model, which is a more general model that fits the sample data perfectly (Singer & Willett, 2003). The deviance is defined as

$$\textit{Deviance} = -2\left[l\left(\tilde{\theta}\right) - l\left(\hat{\theta}\right)\right]. \tag{3.6.5}$$

Since the LL of the saturated model is exactly one, the deviance can be expressed simply as

$$\textit{Deviance} = -2l\left(\tilde{\theta}\right). \tag{3.6.6}$$

According to Singer & Willett (2003) and Hosmer & Lemeshow (2000), the smaller the deviance statistics, the better the fit of a model.

The Akaike information criterion

The Akaike information criterion (AIC) is used to compare the relative goodness-of-fit of competing non-nested models(Hosmer & Lemeshow, 2000). It measures both how well the model fits the data, and how complex the model is. It uses the parsimony standard that says that a model using fewer parameters and explaining the context almost in a same level is best. It is based on the LL statistic, but instead of using the LL itself, it penalises the LL according to pre-specified criteria. The AIC penalty is based on the number of model parameters. This is because addition of parameters, even if they have no effect, increases the LL statistic, resulting in a decrease in the deviance statistics(Box-Steffensmeier & Jones, 2004). The AIC statistic is defined as:

$$\textit{AIC} = -2\left[l\left(\tilde{\theta}\right) - 2p\right], \tag{3.6.7}$$

where p is the number of model parameters.

Bayesian information criterion

Bayesian information criterion (BIC) is another model selection criterion based on information theory, but set within the Bayesian context (Cleves, 2008). It is also used to compare the relative goodness-of-fit of competing non-nested models. However, unlike the AIC, its penalty

is not just based on the number of model parameters, but also on the sample size(Cleves, 2008). The BIC statistic is computed as:

$$BIC = -2\left[l\left(\tilde{\theta}\right) - p\log n\right], \tag{3.6.8}$$

where p is the number of model parameters and n is the sample size. The best model is the one that provides minimum BIC.

3.6.2 Residuals and goodness-of-fit

In Section 3.3.1, maximum likelihood estimation of discrete-time hazard models was based on the binary representation of transitions between states. However, when the interest is in analysis of residuals and goodness-of-fit, one should take into account that the original data consist of n independent observations (t_i, δ_i, x_i). The deviances and residuals obtained from fitting a binary model are not appropriate for discrete-time survival data. This section presents alternative strategies for the construction of valid residuals for discrete-time survival data.

Martingale residuals

The martingale residual takes censoring into account and is particularly suited for assessing the functional forms of predictor effects (Tutz & Schmid, 2016). The martingale residual is defined by:

$$m_i = \delta_i - \sum_{j=1}^{j_i} \hat{h}_{ij}, \quad i = 1, \ldots, n,$$

where $\hat{h}_{ij} = \hat{h}(j \mid x_i)$. We also have $\hat{H}(j_i) = \sum_{j=1}^{j_i} \hat{h}_{ij}$ as the cumulative risk of observation i up to time j_i. The idea behind the martingale residual is to compare for each observation the observed number of events up to j_i (measured by δ_i) with the expected number of events up to j_i (measured by $\hat{H}(j_i)$). Using the binary variables representation with $(y_{i1}, \ldots, y_{ij_i}) = (0, \ldots, \delta_i)$, the residuals can be defined as

$$m_i = \sum_{j=1}^{j_i} \left(y_{ij} - \hat{h}_{ij}\right), \quad i = 1, \ldots, n.$$

For a well-fitting model that includes all relevant explanatory variables, the martingale residuals should be random and uncorrelated with the covariate values (Tutz & Schmid, 2016). To assess the importance of a covariate graphically in a discrete-time survival model, the residuals can be plotted against the covariate values.

Deviance residuals

The deviance residuals are martingale residuals that have been transformed to be more symmetric about zero. They were first introduced by Therneau et al. (1990). Deviance residuals

of a discrete-time survival model can be used to examine the fit of a model on a case-by-case manner. The deviance residuals are also usefull for identifying outliers. The deviance residual for each individual i at time j is calculated via the following formula:For a discrete-time model, the deviance residuals are given by

$$\begin{aligned} r^2_{D,i} &= -2\left[\delta_i \log\left(\hat{P}(T_i = j_i)\right) + (1-\delta_i)\log\left(\hat{P}(T_i > j_i)\right)\right] \\ &= -2\sum_{j=1}^{j_i} y_{ij}\log\left(\hat{h}_{ij}\right) + (1-y_{ij})\log\left(1-\hat{h}_{ij}\right) \end{aligned}$$

The deviance residuals are linked to the familiar maximum likelihood based deviance $D = -2\left[\log\hat{L}_c - \log\hat{L}_N\right]$, such that $D = \sum r^2_{D,i}$.

3.7 Chapter summary

This Chapter provided a theoretical background for the discrete-time single risk and competing risk models. A frailty model that accounts for unobserved heterogeneity in the single risk case was also presented. The distribution of the frailty term was assumed to be Gaussian following the conclusions from Chapter 2. The process for creating the person period data set that is required for fitting the discrete-time single risk and competing risk models was also outlined. The Chapter concluded with a presentation of Goodness-of-fit tests used to assess the adequacy of the fitted discrete-time models. The results of the fitted models as well as the different Goodness-of-fit tests are presented in the next Chapter.

Chapter 4

Data Analysis

4.1 Introduction

This chapter presents an analysis of the data used in the study as well as the study findings. The chapter begins with a description of the data used followed by a presentation of the study results in line with the study objectives. The purpose of the study was to analyse the temporal nature of the process of student dropout using discrete-time survival analysis methods. In order to achieve this purpose, the following main objectives were identified:

- To analyse the incident of dropout. Section 4.2 presents the distribution of each of the variables used in the study. The descriptive statistics reported for the descriptor and predictor variables include the mean, standard deviation and percentages. Baseline survival functions are reported in Section 4.3.1. The probability of dropout in each year is also reported in Section 4.4.1.

- To identify the determinants of dropout. The results of the single risk model based on five years are reported in Section 4.4.1. The determinants of dropout identified through the competing risks model based on Year 3, 4 and 5 data are reported in Section 4.4.2. The estimated coefficients (log odds of dropout), the standard error as well as the p-values are reported for both the single and competing risks model.

- To compare the risk profile of dropping out among different groups of students. The risk of dropout is first compared graphically through survivalr functions presented in Section 4.3.1. The results of the log-rank test for gender, race, residence type and language are presented in Section 4.3.3. The risk profile of dropout is further analysed based on the single and competing risks model estimates.

- To compare the discrete-time single risk model versus the competing risk model. The results are provided in Section 4.4.3.

- To compare the discrete-time single risk model with unobserved heterogeneity versus the discrete-time single risk model without unobserved heterogeneity. The results are provided in Section 4.5.

4.2 Data description and exploratory data analysis

4.2.1 Data description

The data used for the study was obtained from the Tshwane University of Technology (TUT) Intergrated Tertiary System (ITS). The data covers 565 students enrolled for the first year of engineering three year diplomas for the first time in 2010. Students with foreign Matric as well as students who received conditional exemption were excluded from the study as the admission point score (APS) could not be computed for these students. Furthermore, students who did not have Matric mathematics marks were also excluded resulting in a final data set of 502 students. The race variable was collapsed into two categories (White and Non-white) due to the small sample size for Indians (8) and Coloureds(3). This was done to avoid under or over-estimating the regression coefficients. Data was analysed using Stata version 16 statistical analysis software (?).

The cohort was followed for five years from 2010 through 2014, inclusive. In the study the term dropout refers to the act of dropping out of an engineering programme before graduation. The dependent variable is the maximum observation time for each student, i.e., time to dropout, graduation or the last year of the study, if the student was still enrolled at that time.

Demographic and academic variables known to be associated with student academic outcomes that were available on the database were used as explanatory variables. Furthermore, some variables were only used for descriptive purpose. Some potentially important explanatory variables could not be obtained from the database, e.g., financial support, school quintile, first generation student status. The full list of the variables used in the study are presented in 4.1.

Table 4.1: Variables used in the study.

Variable	Description
Gender	Male Female
Race	White Non-white
Residence	On-campus Off-campus Private
English language	First languange Second language
APS	Used as a measure of matric performance
Mathematics score	Used as a measure of performance in Mathematics

4.2.2 Frequency tables and summary statistics

Qualitative descriptive information for the sample of students used in the study is provided in Table 4.2. A summary of continuous variables is presented in Table 4.3. The frequency table presented in Table 4.2 shows that Building Science had the highest number of registered first year engineering students in 2010, i.e. 81 (16.14%). This was followed by Electrical and Civil engineering with 70 (13.94%) and 67 (13.35%) students respectively. The remaining programmes, i.e. Metallurgy, Chemical, Surveying, Industrial, Mechanical and Mechatronics each respectively accounted for 12.15%, 11.35%, 11.35%, 7.97%, 7.57%, and 6.18% of the first year engineering student population in 2010. In terms of gender, an overwhelming majority of the 2010 first year engineering population were males (70.52%) with females representing only 29.48% of the population. The sample is also predominantly non-white, i.e. 77.49%.

Table 4.2: Frequency distribution of qualitative explanatory variables.

Variable	Programme	Frequency	Percent
Qualification	Building Science	86	16.76
	Chemical	57	11.11
	Civil	70	13.65
	Electrical	71	13.84
	Industrial	37	7.21
	Mechanical	43	8.38
	Mechatronics	32	6.24
	Metallurgy	62	12.09
	Surveying	55	10.72
Gender	Female	148	29.48
	Male	354	70.52
English language	First language	114	22.22
	Second language	399	77.78
Residence	On-campus	66	13.15
	Off-campus	90	17.93
	Private	346	68.92
Race	White	113	22.51
	Non-white	389	77.49

In addition, a large proportion of the sampled students, i.e. 68.92% used private accommodation, while 17.93% and 13.15% respectively used off-campus and on-campus University accommodation. Only 22.22% of the sampled students took English as a first language in Matric, the remaining 78.78% took English as second language in Matric.

A summary of the quantitative attributes of the sampled population by gender is provided in Table 4.3. The summary table indicates the average age of both groups was almost equal at 19.55 years and 20.47 years respectively for females and males. We observe from Table 4.3 that the age of the youngest student in the sample was the same for both gender groups.

However, we also note that while the youngest female student was 26 years old the oldest males student was 11 years older at 37 years.

Table 4.3: Summary of quantitative variables by gender.

Variable	Gender	n	Mean	Std dev	Min	Median	Max
Age	Females	148	19.55	1.42	17	19	26
	Males	354	20.47	2.27	17	20	37
APS	Females	148	28.42	3.68	21	28	42
	Males	354	27.68	3.95	12	28	38
Mathematics score	Females	148	4.84	0.99	3	5	7
	Males	354	5.07	1.13	2	5	7

In terms of APS, Table 4.3 indicates that the average APS for females (28.42 points) was slightly higher higher than that of males (27.68 points). On the other hand, the median APS for both groups was equal at 28 points indicating that 50% of males and females had an APS less than or equal to 28 points. This suggest that Matric performance does not differ by gender. We also see from Table 4.3 that the lowest APS, i.e. 12 points was obtained by a male student while the highest, i.e. 42 points was obtained by a female student.

When we look at Matric Mathematics score, we see that the average score was 4.84 points for females and 5.07 points for males. Similar to the APS, the median Matric Mathematics score was equal at 5 points. This indicates that at least 50% of both males and females achieved 60% or more in Mathematics. The smallest score obtained by a female student was 2 points, while 3 points was the smallest score obtained by a female students. The highest score obtained by both gender groups was equal at 7 points.

Table 4.4 depicts a distribution of quantitative variables used in the study by race. We see that the average age at first time enrollment into an engineering programmes in 2010 was about 21 years for whites and 20 years for non-whites. The youngest student in the cohort was non-white and 17 year old. On the other hand, the oldest student was white and 37 years old .

Table 4.4: Summary of quantitative variables by race.

Variable	Race	n	Mean	Std dev	Min	Median	Max
Age	White	113	20.70	2.52	19	20	37
	Non-white	389	20.06	1.93	17	20	32
APS	White	113	28.70	4.09	16	29	37
	Non-white	389	27.67	3.80	12	27	42
Mathematics score	White	113	4.73	1.13	2	5	7
	Non-white	389	5.08	1.07	2	5	7

Table 4.4 further shows that the difference in the average APS between whites and non-whites was 1 point. We also see that at least 50% of white students obtained 29 APS points or more while at least 50% of non-white students obtained 27 APS points or more. There seems to be differences in the minimum and maximum APS points obtained by the two population groups. Specifically, the lowest APS points of 12 was for a non-white student whereas the lowest APS points obtained by a white student was 16 points.

The average performance in Matric Mathematics was not different between the two race groups. For example, the average score for whites was 4.73 points and 5.08 points for non-whites. The minimum, median and maximum scores for the two groups were equal, i.e. 2 points, 5 points and 7 points respectively.

4.2.3 Incidence of dropout

Table 4.5 provides details on the enrollment status of the 502 students from January 2010 till December 2014. From Table 4.5, we see that from the initial 502 students, 58 (11.55%) students had dropped out by the end of the first year, while 41 (9.23%) dropped out in the second year. The dropout rate increased to 15.14% (61) in the third year, decreased to 12.25% (36) in the fourth year and increased again to 19.07% (33) in the fifth year. Overall, a total of 160 and 229 students had dropped out by the end of the third and fifth year respectively , i.e. 31.87% and 45.62%. The dropout rate was the highest in the third (15.14%) and fifth (19.07%) year and lowest in the second (9.23%) and first (11.55%) year.
Table 4.5 further shows that only 48 students completed their studies by the end of the third year, i.e. 9.56%. Completion figures improved to 85 in the fourth year and 82 in the fifth year. Cumulatively, 42.82% (215) of the students completed their studies within five years of registration.

Table 4.5: Enrollment status by gender.

Year	Gender	Enrolled		Dropped out		Graduated	
		Count	Percent	Count	Percent	Count	Percent
2010	Female	148	29.48	12	8.11	0	0.00
	Male	354	70.52	46	12.99	0	0.00
Totall		502	100.00	58	11.55	0	0.00
2011	Female	136	30.63	12	8.82	0	0.00
	Male	308	69.37	29	9.42	0	0.00
Total		444	100.00	41	9.23	0	0.00
2012	Female	124	30.77	22	17.74	14	11.29
	Male	279	69.23	39	13.98	34	12.19
Total		403	100.00	61	15.14	48	11.91
After 3 years	Female			46	31.08	14	9.46
	Male			114	32.20	34	9.60
Total				160	31.87	48	9.56
2013	Female	88	29.93	8	9.09	21	23.86
	Male	206	70.07	28	13.59	64	31.07
Total		294	100.00	36	12.25	85	28.91
2014	Female	59	34.10	14	23.73	26	44.07
	Male	114	65.90	19	16.67	56	49.12
Total		173	100.00	33	19.07	82	47.40
After 5 years	Female			68	45.95	61	41.22
	Male			161	45.48	154	43.50
Total				229	45.62	215	42.83

In terms of gender, the results as shown in Table 4.5 indicate that 12 (8.11%) female students dropped out in the first year of registration compared to 46 (12.99%) male students. The dropout rate continued to be low for female students compared to males in the second year of registration, i.e. 8.82% for females and 9.42% for males. The trend changed in the third, fourth and fifth year of registration with more females dropping out in comparison to males. For instance, 17.74% of females dropped out in the third year of registration compared to 13.98% of males, while 13.59% of females dropped in the fourth year in comparison with 9.09% of males. In the fifth year, the dropout rate for females increased to 23.73% compared to 16.67% for males.

The dropout rate for both groups was the highest in the fifth year, i.e. 23.73% for females and 16.67% for males. The lowest dropout rate for females was recorded in the first year at 8.11% , while fewer males dropped out in the second year compared to all the other years, i.e. 9.09%.

Enrollment status of the 502 students was also summarised in terms of race to see if there are any racial differences emerging. Table 4.6 provides a summary of the dropout rates as

well as graduation rates by race.

Table 4.6: Enrollment status by race.

Year	Race	Enrolled		Dropped out		Graduated	
		Count	Percent	Count	Percent	Count	Percent
2010	White	113	22.51	14	12.39	0	0.00
	Non-white	389	77.49	44	11.31	0	0.00
Total		502	100.00	58	11.55	0	0.00
2011	White	99	22.30	5	5.05	0	0.00
	Non-white	345	77.70	36	10.43	0	0.00
Total		444	100.00	41	9.23	0	0.00
2012	White	94	23.32	7	7.45	21	22.34
	Non-white	309	76.68	54	17.48	27	8.74
Total		403	100.00	61	15.14	48	11.91
After 3 years	White			26	23.00	21	18.58
	Non-white			134	34.45	27	6.94
Total				160	31.87	48	9.56
2013	White	66	22.45	7	10.61	26	39.39
	Non-white	228	77.55	29	12.72	59	25.88
Total		294	100.00	36	12.25	85	29.91
2014	White	33	19.07	3	9.09	24	72.73
	Non-white	140	80.93	30	16.67	58	41.43
Total		173	100.00	33	19.07	82	47.40
After 5 years	White			36	31.86	71	62.28
	Non-white			193	49.61	144	37.02
Total				229	45.62	215	42.83

From Table 4.6 we see that in 2010 the dropout rates of white students was higher than that of non-white students by only one unit, i.e. 12.39% for whites and 11.31% for non-whites. The picture changes considerably in the second and third years with the white population group experiencing a significant decline and non-whites experiencing a slight decline followed by a sharp increase. More specifically, in the second and third year, the dropout rates for white students declined respectively to 5.05% and 7.45% while that of non-white students declined to 10.43% in the second year and increased to 17.48% in the third year.

From Table 4.6 we also see that the dropout rate of white students peaked in the first (12.39%) and fourth (10.61) year, while most non-white students dropped out in the third (17.48%) and fifth (12.72%) year. The dropout rate for non-white students was the lowest in the second (10.43%) and first (11.31%) year, while fewer white students dropped in the second (5.05%) and third (7.45%) year. We also see that 23% of white students had dropped out by the end of the third year compared to 34.45% of non-white students. Similarly, the dropout rate after five years for non-white students is considerably higher than that of white students, i.e.

49.61% versus 31.86%.

4.3 Nonparametric analysis

This section looks at dropout survival probabilities without making assumptions about the functional form of the survivor functions. Visual techniques are used to compare the dropout survival rates by gender, race, residence type and English language. The visual results are further confirmed through the use of formal test of hypothesis for the equality of survivor functions across the groups. The median survival times for the entire sample of students as well as per group are also presented.

4.3.1 Survival functions

Gender

Figure 4.1 presents the baseline survival plot by gender. The curves suggest that female students were likely to survive longer than their male counterparts. However, the difference in the survival probabilities between the two groups does not seem to be much. For instance, about 92% of female students survive past the first year compared to 87% of males students. The survival probability seems to level of by the end of the third year to 0.69 for females and 0.68 for males. The survival probability stepwise declines to 0.48 and 0.49 for females and males respectively in the fifth year.

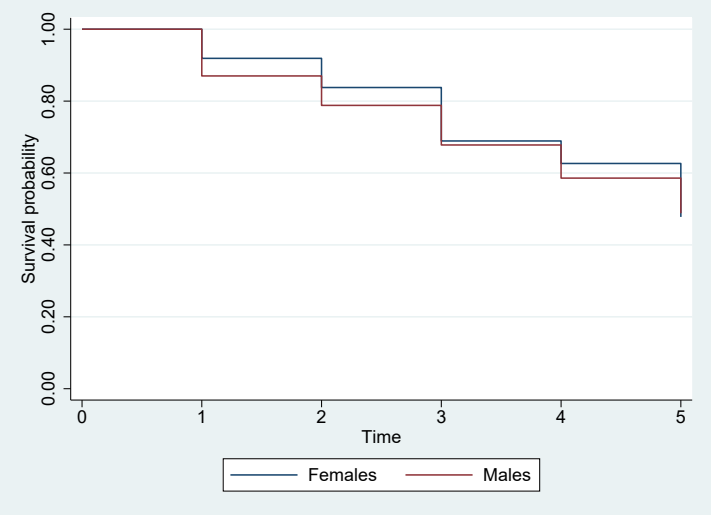

Figure 4.1: Gender baseline survival function

Race

The baseline racial dropout survival probabilities are given as a function of time in Figure 4.2. The dropout pattern for whites and non-whites seems to be similar for the first three years. The first year survival probability is about 88% for whites and 89% for non-whites. The

probability of surviving past the second and third years drops to 0.83 and 0.77 respectively for whites, and 0.79 and 0.66 respectively for non-whites. More whites survive till the fourth year in comparison to non-whites. The difference is about 12 percentage points, i.e. 69% for whites against 57% for non-whites . The difference in survival probabilities between the two groups increases to about 19% points in the fifth year, such that 63% of whites survive till the fifth year compared to about 45% of white students.

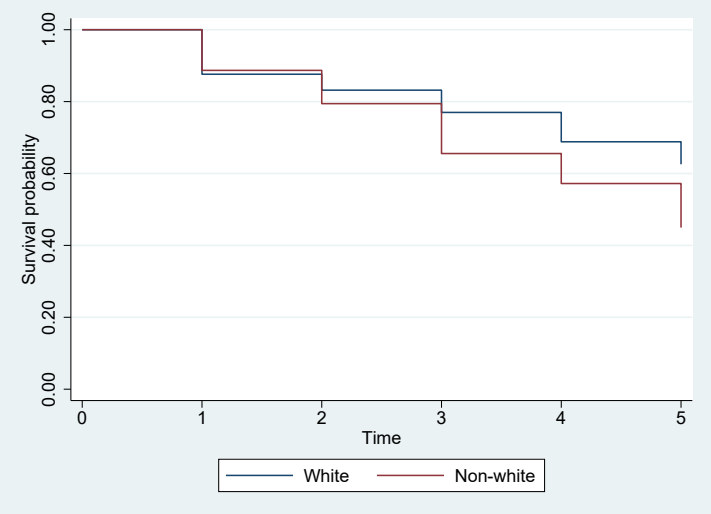

Figure 4.2: Race baseline survival function.

English language

The baseline English language dropout survival probabilities shown in Figure 4.3 indicate that survival probabilities between the two groups are neck on neck in the first, second and third year. The proportion of second language students who survive till the fourth and the fifth year is slightly higher than that of first language students., i.e. about 56% for first language students and 61% for second language students in the fourth year, and about 42% for first language students and 49% for second language students in the fifth year. The results suggest that more second language students survived till the fifth year compared to first language students.

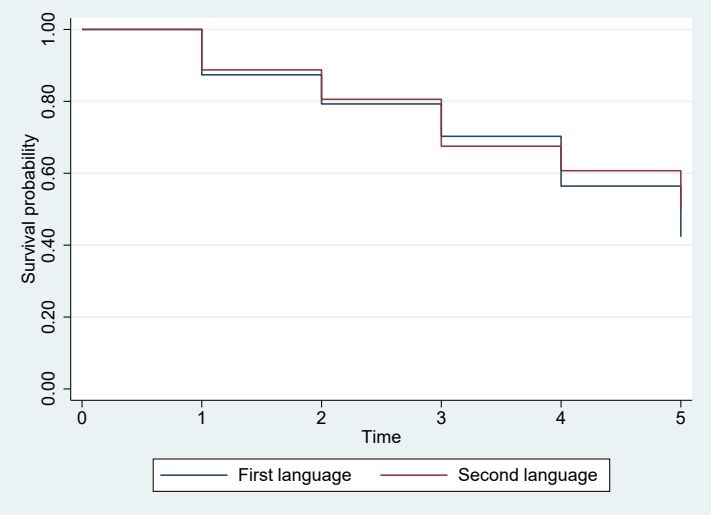

Figure 4.3: Language baseline survival function.

Type of residence

The baseline dropout survival rates by residence type are depicted in Figure 4.4. It is important to note that survival rates are produced for only the first, second and third years as very few students (8) remain on University residences after the third year. Most of the sampled students who survive past the third year use private accommodation. Figure 4.4 suggest that students residing in both on-campus and off-campus based TUT residences are more likely to survive the first year compared to those residing in private accommodation. In particular, we see that 97% of students residing in on-campus based TUT residences survived the first year compared to 92% of those residing in off-campus based TUT residences and about 87% of those residing in private accommodation.

This trend continues into the second year with dropout survival rates of 89%, 82% and 78% for on-campus accommodation based students, off-campus accommodation based students and private accommodation based students, respectively. The results in 4.4 indicate a shift in the survival probabilities in the third year. Specifically, the survival rate for those residing in off-campus based accommodation declines sharply from 82% to about 35%. The probability of surviving till the third year declines to 0.74 for on-campus based students and 0.68 for private accommodation.

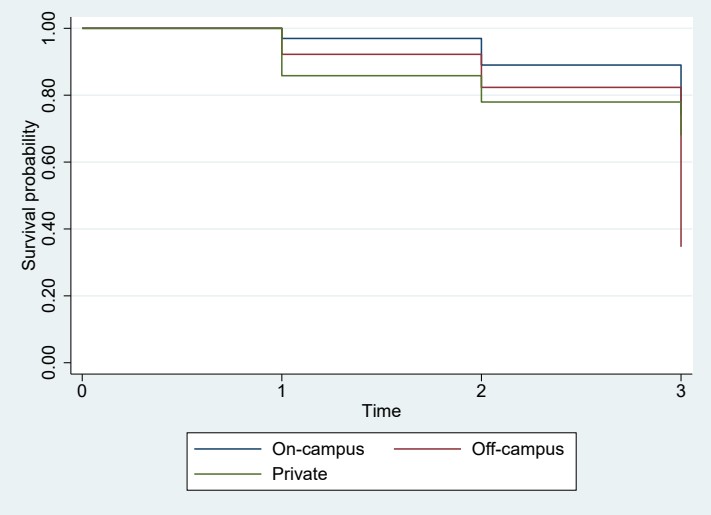

Figure 4.4: Type of residence baseline survival function.

4.3.2 Median survival times

The estimated median lifetime for the entire sample of students and for the specific subgroups is shown in Table 4.7. As shown in Table 4.7, 50% of the sampled students survived till the fifth year. Furthermore, the median survival time does not differ by gender and language. Students residing in on-campus accommodation seem to survive longer than those in private accommodation and off-campus based accommodation. The missing median survival time for on-campus based accommodation indicates that the median survival time is more than five years. We also see that 50% of students residing in private accommodation survive till the fifth year while the same proportion of those residing in off-campus based accommodation survived till the third year.

Table 4.7: Estimated median survival times

Variable	Category	Estimated median survival time
Gender	Male	5
	Female	5
Race	White	-
	Non-white	5
Accomodation	On-campus	-
	Off-campus	3
	Private	5
English	First languange	5
	Second language	5
Overall		5

- Missing

4.3.3 Testing equality of survivor functions

The log-rank test was performed to confirm the graphical results observed in the comparison of dropout survival rates by gender, race, residencee type and English language. The results of the test for each of the variables are presented in Table4.8. The results indicate that the dropout survival probability does not differ significantly by gender, residence type and English language. The results for the comparison on the basis of race indicate that there is a significant difference between survival curves for race.

Table 4.8: Log-rank test for equality of survivor functions.

Variable	Dropout		
	χ^2	DF	p-value
Gender	0.14	1	0.706
Race	6.37	1	0.012 *
English	0.50	1	0.459
Accomodation type	3.91	2	0.142

* $p<0.05$

4.4 Model results

4.4.1 Single risk model

The plots of the estimated hazard and survival probabilities provides information on the timing of dropout, however, they do not give an indication of which students are more at risk of dropping out. In this section the single risk discrete-time hazard model is fitted using logistic regression to determine the effects of covariates on student dropout. This allows for prediction of the risk of dropout based on a set of explanatory variables. We first fit a model that looks at the main effect of time. The model is fitted by including only time point intercepts for the years 1 to 5. Starting the analysis with an initial time-only hazard model provides direct information on the shape of the entire student sample hazard function. The section continues with investigation of effects of explanatory variables on the hazard probabilities by adding explanatory variables to the initial model.

Stepwise regression procedures are not used for variable selection, but rather inclusion of explanatory variables in the model is driven by theory as recommended by (Mills, 2011). A general specification of time using a separate dummy variable for each time period is employed as we only have a few discrete time points, i.e. 5 points and a few explanatory variables. The use of dummy variables representation for the time periods in the model is recommended for studies with few discrete-time points as it does not place any constraint on the shape of the baseline model and it simplifies interpretation of coefficients (Singer & Willett, 2003).

Baseline hazard model

The estimates of the coefficients β, the corresponding estimated standard errors (SE), and the p-values of the time effects obtained from the baseline single risk model are presented in Table 4.9.

Table 4.9: Baseline profile of dropout risk over time.

Variable	Year	β	SE	p-value
Period	1	-2.03	0.14	0.000^{***}
	2	-2.29	0.16	0.000^{***}
	3	-1.72	0.14	0.000^{***}
	4	-1.97	0.18	0.000^{***}
	5	-1.44	0.19	0.000^{***}

*** p<0.001

The results from Table 4.9 indicate that, assuming that the sampled students are homogeneous, i.e. they are not distinguished by values of any explanatory variables, the risk of dropping out in the first year is estimated as almost 12%. To obtain this value we note that from Equation 3.3.16 in section 3.3.1 the conditional probability or the hazard of dropout for each year can be obtained as

$$\hat{h}_j = \frac{1}{1+e^{-(\alpha_l)}}.$$

This implies that the hazard of dropout in the first year is given by

$$\hat{h}_j = \frac{1}{1+e^{-(-2.04)}} = 0.12.$$

Similarly, the risk of dropping out is approximately 9% in the second year, 15% in the third year and 12% in the fourth year. The risk increases to about 19% in the fifth year. The hazard of dropout plotted as a function of time is depicted in Figure 4.5.

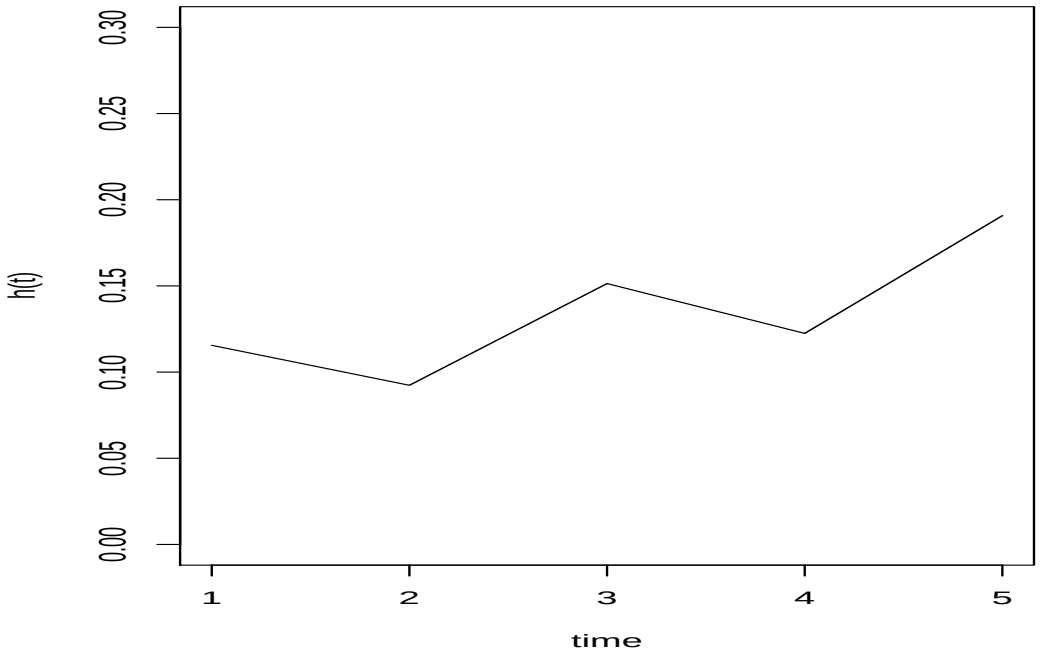

Figure 4.5: Discrete-time hazard function for dropout.

Single risk full model

The second and final model fitted under the single risk specification is the discrete-time logit model adjusting for all covariates. Gender and race are treated as time-invariant and their effects are also assumed to be time-invariant. Residence type is included as a time-varying covariate since students may change their residence type yearly. All the variables that are related to Matric, i.e. APS and language are allowed to interact with time so as to determine whether their effects on dropout vary with time as suggested by literature, even though they are time-invariant. Table 4.10 shows the estimates of the coefficients (β), the corresponding estimated standard errors (SE), and the p-values of the covariate effects obtained from the single risk model.

The results as depicted in Table 4.10 indicate that race has a significant effect on dropout. In particular, we see that the odds of dropout of non-white students is 1.86 times more than that of white-students. The results further show that the effect of gender on dropout is not significant. In terms of language, the results indicate that language affects the risk of dropout significantly only in the fourth year. Specifically, the odds of dropout in the fourth year is 2.32 times higher for students with English as a first language in Matric in comparison to those with English as a second language.

Table 4.10: Maximum likelihood estimates: single risk model.

Variable	Reference group	Year	β	SE	p-value
Gender					
	Male		0.15	0.17	0.304
Race					
	Non-white		0.62	0.21	0.004**
English					
		1	-0.24	0.34	0.713
		2	-0.13	0.40	0.750
	First language	3	-0.42	0.38	0.264
		4	0.84	0.39	0.032*
		5	0.39	0.53	0.455
Residence					
		1	1.01	0.82	0.219
	Off-campus	2	0.36	0.65	0.577
		3	1.88	0.73	0.010*
		1	1.73	0.74	0.019*
	Private	2	0.17	0.51	0.743
		3	-0.20	0.58	0.731
APS					
		1	0.00	0.05	0.949
		2	-0.04	0.05	0.485
		3	-0.06	0.05	0.217
		4	0.03	0.06	0.559
		5	0.06	0.06	0.338
Mathematics score					
		1	-0.26	0.16	0.114
		2	-0.29	0.19	0.129
		3	-0.08	0.16	0.608
		4	0.28	0.22	0.559
		5	-0.42	0.23	0.069
Period					
		2	2.31	1.84	0.208
		3	2.74	1.79	0.125
	Year 1	4	-2.26	1.83	0.300
		5	1.11	1.28	0.578

* $p<0.05$
** $p<0.01$

The results further show that the effect of residence type on dropout is significant only in the first and third years. In the first year, students with private based accommodation are 5.64 times more likely to dropout compared to those residing on campus. However, in the third year the odds of dropout is 6.55 times higher for students residing in off-campus based accommodation compared to on-campus based students. From the results in Table4.10, we also see that the effect of APS on dropout is not significant. Matric Mathematics score also does not have a significant effect on the risk of dropout. Furthermore, the results indicate that when the risk of dropout in the second, third, fourth and fifth year is compared to the risk in the first year, the differences are insignificant.

4.4.2 Competing risk model

In this section dropout and graduation are jointly estimated so as to account for the possible inter correlation between dropout and graduation. The model is estimated for the third, fourth and fifth years since completion only occurs from the third year. The results obtained when dropout is treated as a competing risk are not presented as the focus of the study is on dropout.

Table 4.11: Maximum likelihood estimates: competing risks model

Variable	Reference group	Year	β	SE	p-value
Gender					
	Male		-0.24	0.23	0.917
Race					
	Non-white		0.39	0.30	0.206
English					
	First language	3	-0.49	0.38	0.194
		4	1.35	0.42	0.001**
		5	0.39	0.60	0.518
Mathematics					
		3	-0.34	0.16	0.811
		4	0.37	0.22	0.102
		5	-0.43	0.26	0.090
APS score					
		3	-0.73	0.05	0.122
		4	0.03	0.06	0.566
		5	0.09	0.07	0.240
Period					
	Year 3	4	-5.50	1.88	0.003**
		5	-1.66	2.11	0.432

** $p<0.01$

The competing risk model results are presented in Table 4.11 in terms of the estimated β coefficients, the corresponding estimated standard errors (SE), and the p-values. The results indicate that when graduation is treated as a competing risk, only language and year of study have a significant effect on dropout. For language, we see that the odds of dropout in the fourth year is 3.9 times more for students with English as a first language students compared to second language students. In terms of year of study, the results show that the odds of dropout in the third year is 0.41% less than in the fourth year.

4.4.3 Comparison of single risk and competing risk models

The results of both the single risk discrete-time model formulation and the discrete-time competing risks formulation are presented in Table 4.12 in terms of the estimated β coefficients, the corresponding estimated standard errors (SE), and the p-values. The single risk model results presented are for the model fitted using data for only the third, fourth and fifth years to allow for comparison with the competing risk model fitted when graduation is treated as a competing event. The general pattern of the estimated effects of most of the explanatory variables in the two specifications is similar. However, some of the effects are slightly more pronounced in the competing risk case. For instance the effect of gender on dropout in any year (third, fourth or fifth) are insignificant in both models. APS and mathematics score also do not have a significant effect on dropout in the third, fourth and fifth year in both models.

Table 4.12: Model comparison: single risk versus competing risk.

Variable	Reference group	Year	Single risk			Competing risk		
			β	SE	p-value	β	SE	p-value
Gender								
	Male		-0.05	0.23	0.820	-0.24	0.23	0.917
Race								
	Non-white		0.62	0.30	0.040*	0.39	0.31	0.206
English								
		3	-0.42	0.38	0.265	-0.49	0.38	0.194
	First language	4	0.85	0.34	0.029*	1.35	0.42	0.001**
		5	0.42	0.53	0.423	0.38	0.60	0.518
Mathematics								
		3	-0.06	0.16	0.697	-0.38	0.16	0.811
		4	0.30	0.22	0.165	0.37	0.22	0.102
		5	-0.60	0.23	0.090	-0.43	0.26	0.090
APS								
		3	-0.65	0.05	0.183	-0.73	0.05	0.122
		4	0.30	0.57	0.614	0.03	0.06	0.566
		5	-0.39	0.23	0.390	0.09	0.07	0.240
Period								
	Year 3	4	-4.99	2.20	0.024*	-5.5	1.88	0.003**
		5	-1.60	2.02	0.428	-1.66	2.11	0.432

* $p<0.05$
** $p<0.01$

Similarly, in both models, language has a significant effect on dropout only in the fourth year. However, the results are highly significant in the competing risk model compared to the single risk specification. Furthermore, the risk of dropout of second language English student in the fourth year is 3.9 times higher compared to first language students in the single risk model, while it is 2.4 times higher in the competing risk model.

The racial effect on dropout differs in the two models. The effect in the competing risk case is not significant while it is significant in the single risk case where by the risk of dropout is higher for non-white students compared to white students.

The likelihood of dropout in the third year compared to the fifth year is also significant in both models with the competing risk estimates being highly significant compared to the single risk case. We also see that the risk of dropout is 0.68% lower in the third year compared to the fourth year in the single risk specification while it is 0.41% lower in the competing risk case.

The overall results indicate inconsistencies between the two models for the gender effect.

4.5 Unobserved heterogeneity

The discrete-time single risk model was fitted with unobserved heterogeneity by including a frailty term in the logit model. The frailty term in the model was assumed to have a Gaussian distribution with mean zero. The results of the Likelihood ratio test are depicted in Table 4.13. The σ coefficient reported in Table 4.13 is the standard deviation of the heterogeneity variance. The ρ is the proportion of the total variance contributed by the panel-level variance component. The model chi-square statistic indicate that the regressions are not significant at standard levels. The likelihood ratio tests for ρ indicates a statistically insignificant frailty. The result suggest that the effects of unobserved heterogeneity are insignificant.

Table 4.13: Model comparison: single risk with general time specification versus single risk with linear time specification.

ρ	σ	χ^2	p-value
0.34	1.29	1.02	0.156

4.6 Model adequacy

The discrete-time single risk model fitted based on the general specification of time using indicator variables (general specification) was compared with the model based on time specified as a continuous linear variable to check the fit of the model. The results in Table 4.14 show that the model using indicator variables for fits better. This is based on a BIC of -12178.73 versus -11983.34 and the deviance of 1803 versus 1770.

Table 4.14: Model comparison: single risk with general time specification versus single risk with linear time specification..

Model	Bayesian information criterion	Deviance
General time specification	-11983.43	1770
Linear time specification	-12178.73	1803

Chapter 5

Discussion and conclusions

5.1 Introduction

This chapter presents a summary of the study findings presented in the previous chapter. The conclusions drawn from the findings, the study recommendations as well as limitations are also presented. The purpose of the study the study was to analyse the temporal nature of the student dropout process using discrete-time survival analysis methods. In particular, the main objectives of the study were to: (1) analyse the incidence of dropout, (2) identify determinants of dropout, (2) compare the risk profile of dropout among different groups of students, (4) compare the discrete-time hazard single-risk model with the competing-risk model; and (5) to test the effects of unobserved heterogeneity. The findings of the study are discussed in accordance with the research objectives.

5.2 Incidence of dropout

The results of the study show that 11.55% of the 502 student who registered in 2010 had dropped out by the end of the first year. The dropout rate at the end of the third year was about 32% and 47% by the end of the fifth year. The first year dropout rate is lower than the national dropout rate of 30% reported by Letseka & Breier (2008) for all programmes. The rate is also much lower than the 27% reported for engineering diplomas by Pocock (2012). Overall, dropout was found to be the highest in the third and fifth years, i.e., 15.18% and 19,15% respectively. These, results deviate from the findings from previous studies which indicate that dropout is the highest in the first year.

When looking at the cumulative dropout rate within three years of registration, the dropout rate reported in the study, i.e. 31.87% is lower than the national average of 40% reflected in the 2015 Vital Statistic Public Higher Education for the 2010 cohort of students (CHE, 2015). Similarly, the cumulative five year dropout rate of 47.40% is also lower than the national average of 56% (CHE, 2015).

In terms of graduation, the results indicate that only 9.56% of the cohort completed their

studies within three years, while 42.82% completed their studies within five years. The graduation rate of 9.56% after three years of registration for this cohort is higher than that reported for other UoTs studies. For instance, the CHE 2015 Vital Statistic Public Higher Education report based on the 2010 cohort, reported a graduation rate of 5% within three years of registration CHE (2015). While the five year cumulative graduation rate of 42.8% is slightly lower that the 44% reported by the CHE 2015 Vital Statistic Public Higher Education report CHE (2015), it is much higher than the national figure of 17% reported by Scott et al. (2007) for UoT engineering students. Estimates from the discrete-time single-risk model indicate that the risk of dropout in the first year of registration does not differ significantly to that of subsequent years.

5.3 Determinants of dropout

5.3.1 Gender

In South Africa, females account for a larger share of higher education enrollments than males (CHE, 2010). This is attributed to the fact that fewer females repeat a grade or dropout of school, resulting in more females reaching and passing Matric than males (van Broekhuizen & Spaull, 2017). For instance, in 2018, for every 100 females in Matric there we only 80 males (Spaull & Makaluza, 2019). A study that looked at the 2008 NSC cohort indicate that even though roughly the same number of boys and girls started school in 1997 (49% girls, 51% boys), more females reached Matric than males (van Broekhuizen & Spaull, 2017). According to (Spaull & Makaluza, 2019), females outperform males on average in all subjects and all grades, as well as in the school-leaving exam, i.e. Matric. However, males perform better than females in Mathematics and Physical Science in Matric (Spaull & Makaluza, 2019). This is to a certain extent a function of the higher dropout rates for males in high school, leaving a stronger cohort of males to write Matric (Spaull & Makaluza, 2019). When looking at higher levels of performance, i.e. 60% + males performed better in Mathematics and Physical Science (Spaull & Makaluza, 2019).

However, very few females enroll in engineering programmes (Mangara, 2019). According to Francis (2009) females constitute 53% of students in HE, but tend to cluster in certain fields, specifically Health Sciences and Humanities. The results of the current study show that only 22% of the sampled students were females. This distribution is in line with the distribution at other institutions (Sutherland, 2018; Francis, 2009). When looking at the effect of gender on the risk of dropout, the results of the study indicate that males are 1.15 times more likely to dropout than females. However, this effect is insignificant. The results also show that the percentage of students who had dropped out within three years of registration was almost equal for both genders, i.e., 31.08% for females and 32.20% for of males. These results are also supported by the survival curves for the two groups which showed inconsiderable differences in the survival probability as well as the log-rank test which estimated the median survival time of 5 years for both groups.

These results do not deviate much from the findings of other engineering studies. For instance, Zewotir et al. (2011) found that being female rather than male had no significant effect in failure rates as well as dropout rates of first year engineering students. Francis (2009) also found no significant association between gender and dropout rates of engineering students. However, deviations are observed when the results are compared with findings from the broader HE space.

Bhorat et al. (2012) found that females generally perform better than males in terms of HE throughput and retention. This findings are supported by CHE (2012), which indicate that the course success rates for female students between 2007 and 2012 were consistently between 4 and 5 percentage points higher than they were for males. van Broekhuizen et al. (2016) also found notable gender differences in four year completion rates of the Western Cape Education Department 2006 first-time entering undergraduate cohort. In particular, the results show that four year completion rates where significantly higher for females compared to males, i.e. more than half (52%) of females in the cohort completed an undergraduate qualification by the end of 2009 (van Broekhuizen et al., 2016). In terms of dropout, van Broekhuizen et al. (2016) found that dropout rates within the first three years of registration were marginally lower for females than males.

5.3.2 Language

English is the medium of instruction at most HEIs in South African. Students are therefore, required to study in English. However, most university entrants to South African HEIs institutions are English 2nd language speakers. This is true for the current study where only 22% of the sampled students had English as a first language in Matric. According to Pretorius (2002), many second language students have serious reading comprehension challenges. This in turn means that they have restricted and ineffective access to the rich sources of declarative knowledge provided by print-based material in the learning context (Pretorius, 2002). Language proficiency is regarded as a skill required in academic training, as it can either open the door of academic development, or serve as a barrier(Schaap & Luwes, 2013). According to a CHE report, academic language and language of instruction remain one of the most important barriers to success in HE (CHE, 2010). Vilakazi (2002), asserts that the mastery of the language in which a subject is taught is the prerequisite to the mastery of the subject matter.

Results from studies looking at the relationship between language and success in HE are varied. For instance, Rooney (2015) found that being proficient in English increased the likelihood of graduation. According to Maree (2015), learners taught in their mother tongue consistently outperform those who are not. Moodley (2014) found that learners with English as a first language in Matric are more likely better prepared to deal with the language requirements they will face in HEIs where English is the language of learning and teaching. Moreover, the language proficiency of Matriculants with English as a second language cannot

be regarded as equivalent to the language proficiency of students with English as a first language (Moodley, 2014). Wedekind (2013) found that academic success in the language used for learning at a learner's school is a reliable predictor of academic success in the same language at a HEI. A 2013 TUT study looking at first first year engineering students found that Matric performance in English was to a lesser extent a reliable predictor of success in engineering programmes (Louw, Hofman & Van Wyk, 2013).

On the other hand, some findings suggest that effect of Matric English scores on performance in HE is insignificant. For instance, Venkatas, Rampersad & Mashige (2014), found a weak correlation between Matric English scores and performance in first year optometry modules. In terms of engineering Cliff & Hanslo (2010), found that performance in English on its own is not significantly associated with academic performance for engineering students. The results of the current study differ from the finding of other studies in the sense that they indicate that the effect of language proficiency on the risk of dropout varies with time. In particular, although not significant, the results show that having English as a first language reduced the risk of dropout in the first, second and third year of registration. In the fourth year and fifth year of registration, students with English as first language were more likely to dropout compared to second language students. The effect in the fourth year was found to be significant. The risk of dropout for English first language students was found to be 2.31 times higher than that of second language students. The survivor function show that about 61% of English second language students survived till the fourth year compared to 56% of first language students. The insignificant results could be partly attributed to the fact that first year as well as second and third year engineering modules, although taught in English require other specific understanding of concepts and terminologies.

It should be noted that the study looked at whether English was taken as a first language or a second language. Actual performance in the language was not considered. It can be argued that there could be students who took English as an additional language but performed better than first language students. This would largely be students from good performing schools albeit they are based in disadvantaged communities. Scores from language assessments test would perhaps provide a more reliable measure of proficiency as compared to the first language variable.

Further analysis of the results showed that 83.78% of the 111 students with English as a first language were non-white (results not shown). This suggest that this group was largely made up of non-white students from privileged backgrounds whose parents can afford access to better resourced schools. One can argue that this group would be more likely to dropout compared to their underprivileged counterparts as their privilege affords them access to other opportunities. Dropout for these students would not necessarily be linked to financial reasons.

5.3.3 Matric performance

South African HEIs rely considerably on marks obtained from standardised based school-leaving examinations. Marks obtained from these certificates are treated as indicators of learner's current knowledge as well as their ability to succeed in the HE. The marks are treated as trustworthy signals of ability when comparing students against each other across time, owing to the fact that the school-leaving exams are quality controlled and standardised nationally.

Research on the predictive value of Matric performance for future academic performance points to inconsistencies in the results. Some reports indicate that Matric results have a good predictive value for certain groups of students such as those who received quality high school education and whose first language is English (Essack et al., 2012). Other studies have reported racial Kirby (2013); Foxcroft (2006) and gender Foxcroft (2006) discrepancies. For example, UCT's Alternative Admissions Research Project found that for a cohort of white engineering students, performance in Matric explained significant amounts of variation in academic performance at the end of first year engineering studies. However, performance in Matric explained a much smaller percent of variation in first year academic performance of the cohort of black engineering students enrolled in the same class as the white students.

In other studies,Cliff & Hanslo (2010) found that a weighted APS is significantly associated with academic performance for engineering. On the other hand, (Marnewick, 2012) found no correlation between first year university performance for information technology (IT) students and learners' Matric results. The study by Venkatas et al. (2014) also supports these findings.

The results of the current study show that Matric performance has no significant effect on dropout rate in each year of the five years of the study period. A separate analysis for each gender and race group produced similar results. The effects of Matric performance, albeit insignificant was found to vary with time. For instance, in the first year, a unit increase in APS score was associated with an increase in the dropout risk. However, in the second and third years, a unit increase in APS was associated with a decrease in the risk of dropout. Similarly, a unit increase in APS in the fourth and fifth years was associated with an increase the dropout risk.

5.3.4 Mathematics score

Mathematics is regarded as an essential prerequisite for engineering sciences. It is a vital course in the engineering curriculum. Many aspects of engineering activity require formulating a problem correctly and finding an appropriate method to solve the problem (Steenkamp & Muyengwa, 2018). The need for engineering students to think mathematically and to use mathematics to describe and analyse different aspects of the real world they seek to engineer is widely acknowledged. A study at the University of Pretoria indicate that first year students lack a basic understanding of fundamental mathematical concepts (Steyn et al., 2008).

According to Uysal (2012), inadequate skills in basic mathematics cause problems for students majoring in engineering. Most engineering faculties in South Africa, require a relatively high Mathematics mark and put significant weight on the Mathematics mark as a requirement for admission. TUT requires a pass of NCS Mathematics of at least level 4, and do not admit students who have studied mathematical literacy. At TUT, twenty out of the twenty four subjects in the engineering national diploma are mathematical in nature and use mathematical concepts extensively. One would expect a strong correlation between performance in Matric Mathematics and student performance and consequently dropout.

The results of the study indicate that the effect of Matric Mathematics mark on dropout is not significant in any year. However, the effect was found to be time-varying in the sense that an increase in Mathematics score resulted in a decrease in the risk of dropout in the first, second, third and fifth years of study, and it resulted in an increase in the risk of dropout in the fourth year. The median Mathematics score for the sample was 5 points indicating that at least 50% of the students obtained a score of 5 points or more. Further analysis through a frequency distribution of Mathematics scores indicate that only 33.47% of the students obtained a score below 6 points. This suggest that the cohort consisted mainly of high performers in Mathematics making the group homogeneous with regards to Mathematics scores.

The results of the study are similar to the one reported by (Schaap & Luwes, 2013). Schaap & Luwes (2013) found the correlation between Matric Mathematics score and academic performance in first, second, third and fourth year of university not significant. Schoer et al. (2010) also found Mathematics results in the 2009 NSC examinations to be a not reliable predictor of performance in commerce-related university programmes. However, the results differ to the findings of a study at the North West University which found a significant correlation between NSC Matric Mathematics and Physical sciences mark and performance in the first year of engineering studies (Hattingh, 2011). They also differ with those from an Engineering TUT study which found Matric Mathematics score to be a reliable predictor of success in engineering programmes (Louw et al., 2013).

5.3.5 Accommodation

Studies indicate that students residing in on-campus based accommodation have better academic outcomes as they are more likely to have more time to study. According to Pillay & Ngcobo (2010), accommodation issues was one of the stress factors making progression through to the next year in HE difficult. Zewotir et al. (2011) found that the probability of failure of the first-year engineering students residing in not residence based accommodation is 1.49 times higher than would be the case for someone residing in residence based accommodation. However, in terms of dropout rates, Zewotir et al. (2011) found no significant differences by type of residence. Bengesai & Paideya (2018), analysed the relationship between timely graduation and academic and institutional factors for the 2009, 2010 and 2011 cohorts of

engineering students at UKZN. The results indicate that staying in a university residence was negatively associated with graduation Bengesai & Paideya (2018).

The results of the current study show that type of residence has a time-varying significant effect on the risk of drop out in the first and third years. In Year 1, it was found that students with private based accommodation are 5.64 times likely to dropout compared to those residing on campus. However, the opposite was observed in the third year where students residing in private based accommodation are 0.82 less likely to dropout in comparison to those residing on-campus. In terms of those residing off campus, the results indicate that they were 6.55 times more likely to dropout compared to those residing in university on-campus based accommodation. It is important to note that TUT has a strict performance-based accommodation policy where readmission in the following year is based on academic performance in the preceding year. This results in many students losing their place in University based residences yearly and moving to private based accommodation. Consequently, very few students remain in university based accommodation in year four and year five.

5.3.6 Race

There is a general believe that the continuing racial imbalance in the quality of schooling and in educational outcomes is an important factor behind different success rates for black and white students in HE, and in Engineering in particular. There is strong evidence to suggest that the effect of race on academic outcomes is significant. Bengesai & Paideya (2018), found that African students registered in engineering degrees in 2009, 2010 and 2011 were less likely to graduate in record time compared to the "other" racial groups, where other included, Indian, White and Coloureds. The findings by Sampson (2011) indicate that there is a significant strong association between graduation rates and race, such that African students had the lowest graduation rate followed by the Coloured, Indians and white students. It should be noted that the racial categorisation differs with the one used in the current study where the White population group is a standalone category whereas Coloureds and Indians have been grouped with Blacks. Rooney (2015); Murray (2014) also found race to be a significant determinant of academic success in HE. The findings by Zewotir et al. (2011) also support these results. These results are in line with the findings of the current study which show that non-white students are 1.86 times more likely to dropout compared to white students.

5.4 Model comparison

Comparison of the discrete-time single risk model with the discrete-time competing risk model results revealed inconsistencies between the two models. For instance, the effect of race on the risk of dropout differs in the two models. The effect in the competing risk case is not significant, while it is significant in the single risk case such that the risk of dropout is higher for non-white students compared to white students.

Language has a significant effect on dropout in both models, however, the results are highly significant in the competing risk model compared to the single risk specification. It is also important to note that the risk of dropout of first language English students compared to second language students in the fourth year is 3.9 times higher in the single risk model and 2.4 times higher in the competing risk case.

The effect of gender, APS and Mathematics score are insignificant in both models.

When the discrete-time single risk model without unobserved heterogeneity is compared with the one with unobserved heterogeneity, the results show that the effect of unobserved heterogeneity is not significant.

5.5 Conclusion

The results of the study show the kind of additional insights gained by using discrete-time survival analysis methods to analyse student dropout in comparison to traditional methods like logistic regression. For instance, not only were we able to estimate the risk of dropout in each year, but we were also able to identify the periods of high risk. It was found that the risk was the highest in the fifth year. The inclusion of time-varying variables, in this case residence type, allowed us not to incorrectly treat it as constant, but to rather analyse its effect on the risk of dropout in each year of the five years of the study based on its changing values. The effect was not only found to be significant in some years and insignificant in others, but the effect was also found to differ between the years. The use of a discrete-time model also allowed us to test the effects of APS and Mathematics score on the risk of dropout over the five years. The assumption was that the effect would be more pronounced in the first year of study and then diminish over time. In both cases the effect was found not to be significant in all years, suggesting that the effect is not significant and is constant over time.

The use of the discrete-time competing risk model allowed us to account for the possible correlation between dropout and graduation. Furthermore, inclusion of a random term in the discrete-time model allowed us to test the effects of unobserved heterogeneity.

Recommendations from this study are that discrete-time survival analysis model is more efficient than traditional methods in analysis of the student dropout process and should therefore be used for analysis of academic outcomes such a dropout. The model can account for the temporal nature of the process of dropout. Both time-varying and time-invariant explanatory variables can be included in the model. The effects of time-invariant explanatory variables that might have time-varying effects can also be investigated.

Given the significant effects of race, and type of residence on the risk of dropout, more attention needs to be paid on these variables. More research is needed to unpack underlying issues associated with these variables. The high dropout risk for English first language students compared to second language students in the fourth year needs to be further investigated as

second language students are expected to be more at risk compared to first language students. For this cohort, dropout was not found to be the highest in the first year as suggested by previous research, however, it was found to be the highest in the fifth year. This can possibly be linked to student exclusion rules, which have a bearing on how long a student can linger on in the system without graduating. According to TUT policy, students who failed in passing more than 50% of the prescribed subjects for a particular year of study are excluded. The exclusion can be appealed, the appeals are dealt through on a case-by-case case. The high dropout rate may also be linked to challenges with work integrated learning (WIL) program placement. Further research is needed to understand the role played by WIL in the student retention issues.

The main limitation of this study is based on the use of secondary data. Model estimation is thus limited to the variables that are available on the database. Information on financial support and first generation (FG) status of students was not available on the database even though their effects on academic outcomes, particularly dropout are known (Rooney, 2015; Siyengo, 2015; Moeketsi & Mgutshini, 2014; Murray, 2014; Pocock, 2012; Letseka, 2009). Other variables that measure non-cognitive skills that are known to be important for success in HE such as motivation Sikhwari (2014); Fraser & Killen (2005), self-discipline Fraser & Killen (2005) and engagement Schreiber & Yu (2016); Strydom et al. (2010) were also not available. The other limitation is that censoring is assumed to be random, i.e., non-informative in discrete-time models. However, for students who stay long in the system without graduating or dropping out, this assumption, may or may not be violated due to exclusion rules.

References

Abedi, J. & Benkin, E. (1987). The effects of students' academic, financial, and demographic variables on time to the doctorate [Journal Article]. Research in Higher Education, 27(1), 3-14.

Agresti, A. (1990). Categorical data analysis [Book]. New York: John Wiley and Sons.

Aina, C., Baici, E. & Casalone, G. (2011). Time to degree: students' abilities, university characteristics or something else? evidence from italy [Journal Article]. Education Economics, 19(3), 311-325.

Al-Radaideh, Q. A., Al-Shawakfa, E. M. & Al-Najjar, M. I. (2006). Mining student data using decision trees [Conference Proceedings]. In International arab conference on information technology (acit'2006). Yarmouk University, Jordan.

Alarcon, G. M. & Edwards, J. M. (2013). Ability and motivation: Assessing individual factors that contribute to university retention [Journal Article]. Journal of Educational Psychology, 105(1), 129.

Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. Higher Education Studies, 6(2), 1–18.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories [Journal Article]. Sociological Methodology, 13, 61-98.

Allison, P. D. (1984). Event history analysis: Regression for longitudinal event data [Book]. Beverly Hills, CA: Sage.

Allison, P. D. (2010). Survival analysis using sas:a practical guide [Book]. North Carolina: SAS Institute Inc.

Allison, P. D. (2014). Event history and survival analysis: regression for longitudinal event data (Vol. 46) [Book]. Beverly Hills, CA: Sage.

Ameri, S., Fard, M. J., Chinnam, R. B. & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts [Conference Proceedings]. In Proceedings of the 25th acm international on conference on information and knowledge management (p. 903-912). ACM.

Ampaw, F. D. & Jaeger, A. J. (2012). Completing the three stages of doctoral education: An event history analysis [Journal Article]. Research in Higher Education, 53(6), 640-660. Retrieved from https://doi.org/10.1007/s11162-011-9250-3 doi: 10.1007/s11162-011-9250-3

Ansell, J., Harrison, T. & Archibald, T. (2007). Identifying cross-selling opportunities, using lifestyle segmentation and survival analysis [Journal Article]. Marketing Intelligence and Planning, 25(4), 394-410.

Ashby, A. (2004). Monitoring student retention in the open university: Definition, measurement, interpretation and action [Journal Article]. Open Learning: The Journal of Open, Distance and e-Learning, 19(1), 65–77.

Astin, A. W. (1975). Preventing students from dropping out [Book]. San Francisco: Jossey-Bass.

Baard, R., Steenkamp, L., Frick, B. & Kidd, M. (2010). Factors influencing success in first-year accounting at a south african university: The profile of a successful first-year accounting student. South African Journal of Accounting Research, 24(1), 129–147.

Baird, L. L. (1990). Disciplines and doctorates: The relationships between orogram characteristics and the duration of doctoral study [Journal Article]. Research in Higher Education, 31(4), 369-385.

Baker, M. & Melino, A. (2000). Duration dependence and nonparametric heterogeneity: A monte carlo study [Journal Article]. Journal of Econometrics, 96(2), 357-393.

Barnett, A. G., Batra, R., Graves, N., Edgeworth, J., Robotham, J. & Cooper, B. (2009). Using a longitudinal model to estimate the effect of methicillin-resistant staphylococcus aureus infection on length of stay in an intensive care unit [Journal Article]. American journal of epidemiology, 170(9), 1186-1194.

Barros, C. P., Butler, R. & Correia, A. (2010). The length of stay of golf tourism: A survival analysis [Journal Article]. Tourism Management, 31(1), 13-21.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition [Journal Article]. Research in Higher Education, 12(2), 155–187.

Bean, J. P. & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition [Journal Article]. Review of Educational Research, 55(4), 485–540.

Bengesai, A. V. & Paideya, V. (2018). An analysis of academic and institutional factors affecting graduation among engineering students at a south african university. African Journal of Research in Mathematics, Science and Technology Education, 22(2), 137-148.

Berge, Z. L. & Huang, Y.-P. (2004). A model for sustainable student retention: a holistic perspective on the student dropout problem with special attention to e-learning [Journal Article]. Doesnews(online), 13(5), 501-515.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment [Journal Article]. Journal of the American Statistical Association, 47(259), 501-515.

Bhorat, H., Mayet, N. & Visser, M. (2012). Student graduation, labour market destinations and employment earnings (Tech. Rep.). University of Cape Town, South Africa.

Bilgicer, T., Jedidi, K., Lehmann, D. R. & Neslin, S. A. (2015). Social contagion and customer adoption of new sales channels [Journal Article]. Journal of Retailing, 91(2), 254-271.

Blom, R. (2014). The value of designated subjects in terms of likely student success in higher education (Report). Pretoria.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy [Journal Article]. Journal of the Royal Statistical Society. Series B (Methodological), 11(1), 15-53.

Bogard, M., Helbig, T., Huff, G. & James, C. (2011). A comparison of empirical models for predicting student retention (Report). Western Kentucky University.

Bokaba, K. G. & Tewari, D. D. (2014). Determinats od student success at a south african university:an econometric analysis [Journal Article]. Anthropologist, 17(1), 259-277.

Bowers, A. J. (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts [Journal Article]. The Journal of Educational Research, 103(3), 191-207.

Box-Steffensmeier, J. M. & Jones, B. S. (2004). Event history modeling: A guide for social scientists [Book]. Cambridge University Press.

Breier, M. (2010). From 'financial considerations' to 'poverty': towards a reconceptualisation of the role of finances in higher education student drop out. Higher Education, 60(6), 657–670.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author) [Journal Article]. Statistical Science, 16(3), 199-231.

Breslow, N. (1974). Covariance analysis of censored survival data [Journal Article]. Biometrics, 89-99.

Brown, S. J., Goetzmann, W. N. & Park, J. (2001). Careers and survival: Competition and risk in the hedge fund and cta industry [Journal Article]. The Journal of Finance, 56(5), 1869-1886.

Bruinsma, M. & Jansen, E. P. (2009). When will i succeed in my first-year diploma? survival analysis in dutch higher education [Journal Article]. Higher Education Research and Development, 28(1), 99-114.

Butler, J. S., Anderson, K. H. & Burkhauser, R. V. (1989). Work and health after retirement: a competing risks model with semiparametric unobserved heterogeneity [Journal Article]. The Review of Economics and Statistics, 46-53.

Cabrera, A. F. (1994). Logistic regression analysis in higher education:an applied perspective [Journal Article]. Higher Education: Handbook of Theory and Research, 10, 225-256.

Cabrera, A. F., Nora, A. & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. The Journal of Higher Education, 64(2), 123–139.

Cabrera, A. F., Stampen, J. O. & Hansen, W. L. (1990). Exploring the effects of ability to pay on persistence in college [Journal Article]. The Review of Higher Education, 13(3), 303-336.

Carr, J. C., Haggard, K. S., Hmieleski, K. M. & Zahra, S. A. (2010). A study of the moderating effects of firm age at internationalization on firm survival and short-term growth. Strategic Entrepreneurship Journal, 4(2), 183–192.

CHE. (2010). Access and throughput in south african higher education: Three case studies [Report]. Pretoria, South Africa: Council on Higher Education.

CHE. (2011). Vitalstats: Public higher education 2011 (Report). Pretoria, South Africa: Council on Higher Education.

CHE. (2012). Vitalstats: Public higher education 2012 (Report). Pretoria, South Africa.

CHE. (2013). Vitalstats: Public higher education 2013 (Report). Pretoria, South Africa.

CHE. (2014). Vitalstats: Public higher education 2014 (Report). Pretoria, South Africa.

CHE. (2015). Vitalstats: Public higher education 2015 (Report). Pretoria, South Africa.

Chen, R. & DesJardins, S. L. (2008). Exploring the effects of financial aid on the gap in student dropout risks by income level [Journal Article]. Research in Higher Education, 49(1), 1-18. Retrieved from https://doi.org/10.1007/s11162-007-9060-9 doi: 10.1007/s11162-007-9060-9

Chen, R. & DesJardins, S. L. (2010). Investigating the impact of financial aid on student dropout risks: Racial and ethnic differences [Journal Article]. The Journal of Higher Education, 81(2), 179-208.

Chimka, J. R., Reed-Rhoads, T. & Barker, K. (2007). Proportional hazards models of graduation [Journal Article]. Journal of College Student Retention: Research, Theory and Practice, 9(2), 221-232.

Clarke, P. M., Walter, S. J., Hayen, A., Mallon, W. J., Heijmans, J. & Studdert, D. M. (2012). Survival of the fittest: retrospective cohort study of the longevity of olympic medallists in the modern era [Journal Article]. British Journal of Sports Nedicine, 345, e8308.

Cleves, M. (2008). An introduction to survival analysis using stata [Book]. Stata Press.

Cliff, A. & Hanslo, M. (2010). The design and use of 'alternate'assessments of academic literacy as selection mechanisms in higher education. Southern African Linguistics and Applied Language Studies, 27(3), 265-276.

Cox, D. R. (1972). Regression models and life-tables [Journal Article]. Journal of the Royal Statistical Society, 187-220.

Dekker, G. W., Pechenizkiy, M. & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. [Conference Proceedings]. In Proceedings of the 2nd international conference on educational data mining.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. Decision Support Systems, 49(4), 498-506.

Delen, D. (2011). Predicting student attrition with data mining methods [Journal Article]. Journal of College Student Retention: Research, Theory and Practice, 13(1), 17-35.

De Leonardis, D. & Rocci, R. (2008). Assessing the default risk by means of a discrete-time survival analysis approach [Journal Article]. Applied Stochastic Models in Business and Industry, 24(4), 291-306.

de Rouen Jr, K. R. & Sobek, D. (2004). The dynamics of civil war duration and outcome [Journal Article]. Journal of Peace Research, 41(3), 303-320.

DesJardins, S. L. (2003). Event history methods: Conceptual issues and an application to student departure from college [Book Section]. In Higher education: Handbook of theory and research (p. 421-471). Springer.

DesJardins, S. L., Ahlburg, D. A. & McCall, B. P. (1999). An event history of student departure [Journal Article]. Economic of Education Review, 18(3), 375-390.

DesJardins, S. L., Ahlburg, D. A. & McCall, B. P. (2002). A temporal investigation of factors related to timely degree completion [Journal Article]. The Journal of Higher Education, 73(5), 555-581.

Dey, E. L. & Astin, A. W. (1993). Statistical alternatives for studying student retention:a comparitive analysis of logit,probit and linear regression [Journal Article]. Research In Higher Education, 34(5), 569-581.

Diermeier, D. & Stevenson, R. T. (1999). Cabinet survival and competing risks [Journal Article]. American Journal of Political Science, 1051-1068.

Doyle, W. R. (2009). The effect of community college enrollment on bachelor's degree completion [Journal Article]. Economics of Education Review, 28(2), 199-206.

Du Plessis, S. & Botha, H. (2012). Mining wellness and performance data to identify at-risk first-year engineering students [Conference Proceedings]. In Proceedings of the world congress on engineering (Vol. 1).

Eiselen, R., Jonck, B. & Strauss, J. (2007). A basic mathematical skills test as predictor of performance at tertiary level [Journal Article]. South African Journal of Higher Education, 21(1), 38-49.

Essack, S., Wedekind, V. & Naidoo, I. (2012). Selection tools predictive of success in the health sciences-nated vs. nsc students. South African Journal of Higher Education, 26(3), 472–486.

Fleming, T. R. & Lin, D. (2000). Survival analysis in clinical trials: past developments and future directions [Journal Article]. Biometrics, 56(4), 971-983.

Flinn, C. J. & Heckman, J. J. (1982). New methods for analyzing individual event histories [Journal Article]. Sociological Methodology, 13, 99-140.

Fowler, M. & Luna, G. (2009). High school and college partnerships: Credit-based transition programs [Journal Article]. American Secondary Education, 38, 62–76.

Foxcroft, C. D. (2006). Evaluating the school-leaving examination against measurement principles and methods. In Marking matric: Colloquium proceedings (p. 58-71).

Francis, M. M. (2009). Women in engineering: Identifying and analyzing gender socialization in the faculty of engineering at the university of kwazulu-natal (Unpublished doctoral dissertation). University of Kwazulu-Natal.

Fraser, W. & Killen, R. (2005). The perceptions of students and lecturers of some factors influencing academic performance at two south african universities. Perspectives in Education, 23(1), 25–40.

Geisser, S. (1993). Predictive inference: An introduction. CRC press. Retrieved from https://books.google.co.za/books?id=MKjZnQEACAAJ

Gibbons, R. D., Duan, N., Meltzer, D., Pope, A., Penhoet, E. D., Dubler, N. N., … Henderson, M. (2003). Waiting for organ transplantation: results of an analysis by an institute of medicine committee [Journal Article]. Biostatistics, 4(2), 207-222.

Giot, P. & Schwienbacher, A. (2007). Ipos, trade sales and liquidations: Modelling venture capital exits using survival analysis [Journal Article]. Journal of Banking and Finance, 31(3), 679-702.

Goodman, S., Jaffer, T., Keresztesi, M., Mamdani, F., Mokgatle, D., Musariri, M., … Schlechter, A. (2011). An investigation of the relationship between students' motivation and academic performance as mediated by effort. South African Journal of Psychology, 41(3), 373–385.

Gordon, S. C. (2002). Stochastic dependence in competing risks [Journal Article]. American Journal of Political Science, 200-217.

Gregory-Smith, I., Thompson, S. & Wright, P. W. (2009). Fired or retired? a competing risks analysis of chief executive turnover [Journal Article]. The Economic Journal, 119(536), 463-481.

Gury, N. (2011). Dropping out of higher education in france: a micro-economic approach using survival analysis [Journal Article]. Education Economics, 19(1), 51-64.

Gutierrez, R. G. (2002). Parametric frailty and shared frailty survival models. The Stata Journal, 2(1), 22–44.

Hagedorn, L. S. (2005). How to define retention:a new look at an old problem. Praeger.

Han, J., Kamber, M. & Pei, J. (2006). Data mining: concepts and techniques [Book]. San Francisco: Morgan Kauffman.

Harrison, T. & Ansell, J. (2002). Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities [Journal Article]. Journal of Financial Services Marketing, 6(3), 229-239.

Hattingh, E. (2011). Predicting academic performance in engineering studies. In Conference of the south african society for engineering education (p. 80).

Haybittle, J. (1965). A two-parameter model for the survival curve of treated cancer patients [Journal Article]. Journal of the American Statistical Association, 60(309), 16-26.

Heckman, J. & Singer, B. (1984a). Econometric duration analysis [Journal Article]. Journal of Econometrics, 24(1-2), 63-132.

Heckman, J. & Singer, B. (1984b). A method for minimizing the impact of distributional assumptions in econometric models for duration data [Journal Article]. Econometrica: Journal of the Econometric Society, 271-320.

Henley, W., Rogers, K., Harkins, L. & Wood, J. (2006). A comparison of survival models for assessing risk of racehorse fatality. Preventive Veterinary Medicine, 74(1), 3–20.

Henry, K. L., Thornberry, T. P. & Huizinga, D. H. (2009). A discrete-time survival analysis of the relationship between truancy and the onset of marijuana use [Journal Article]. Journal of Studies on Alcohol and Drugs, 70(1), 5-15.

Hensher, D. A. & Mannering, F. L. (1994). Hazard-based duration models and their application to transport analysis. Transport Reviews, 14(1), 63–82.

Herrera, O. L. (2006). Investigation of the role of pre-and post-admission variables in undergraduate institutional persistence, using a markov student flow model (Thesis).

Hess, W. & Persson, M. (2012). The duration of trade revisited. Empirical Economics, 43(3), 1083–1107.

Heymann, L. & Carolissen, R. (2011). The concept of'first-generation student'in the literature: implications for south african higher education. South African Journal of Higher Education, 25(7), 1378–1396.

Hiemstra, M., Otten, R. & Engels, R. C. M. E. (2012). Smoking onset and the time-varying effects of self-efficacy, environmental smoking, and smoking-specific parenting by using discrete-time survival analysis [Journal Article]. Journal of Behavioral Medicine, 35(2), 240-251. Retrieved from https://doi.org/10.1007/s10865-011-9355-3 doi: 10.1007/s10865-011-9355-3

Hosmer, D. W. & Lemeshow, S. (2000). Applied logistc regression [Book]. New York: John Wiley and Sons.

Hougaard, P. (1995). Frailty models for survival data [Journal Article]. Lifetime Data Analysis, 1(3), 255-273.

Hougaard, P. (2000). Shared frailty models. In Analysis of multivariate survival data (pp. 215–262). Springer.

Hough, G., Garitta, L. & Gómez, G. (2006). Sensory shelf-life predictions by survival analysis accelerated storage models [Journal Article]. Food Quality and Preference, 17(6), 468-473.

Hovdhaugen, E. (2015). Working while studying: the impact of term-time employment on dropout rates. Journal of Education and Work, 28(6), 631–651.

Ingersoll, G. M., Lee, K. L. & Peng, C.-Y. J. (2010). An introduction to logistic regression analysis and reporting [Journal Article]. The Journal of Educational Research, 96(1), 3-14.

Ishitani, T. T. (n.d.). Longitudinal approach to assesing attrition behaviour among first-generation students:time-varinging effects of pre-college characteristics. Research in Higher Education, 44(4).

Ishitani, T. T. (2006). Studying student attrition and degree completion behaviour among first-generation college students in united states [Journal Article]. The Journal of Higher Education, 77(5), 861-885.

Ishitani, T. T. (2008). How do transfers survive after "transfer shock"?a longitudinal study of transfer student departure at a four-year institution. [Journal Article]. Research in Higher Education, 49, 403-419.

Ishitani, T. T. & DesJardins, S. L. (2002). A longitudinal investigation of dropout from college in the united states [Journal Article]. Journal of College Student Retention: Research, Theory and Practice, 4(2), 173-201.

Jacobs, M. (2015). An investigation of the use of nbts in placement of first year students in sciences [Unpublished Work].

Jadrić, M., Garača, e. & Čukušić, M. (2010). Student dropout analysis with application of data mining methods [Journal Article]. Journal of Contemporary Management Issues, 15(1), 31-46.

Jawitz, J. (1995). Performance in first-and second-year engineering at uct. South African Journal of Higher Education, 9(1), 101–108.

Jones, B. S. & Branton, R. P. (2005). Beyond logit and probit: Cox duration models of single, repeating, and competing events for state policy adoption. State Politics and Policy Quarterly, 5(4), 420–443.

Jonson, I. Y. (2006). Analysis of stopout behavior at a public research university: The multi-spell discrete-time approach [Journal Article]. Research in Higher Education, 47(8), 905-934.

Ju, Y., Jeon, S. Y. & Sohn, S. Y. (2015). Behavioral technology credit scoring model with time-dependent covariates for stress test [Journal Article]. European Journal of Operational Research, 242(3), 910-919.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification [Journal Article]. Cybernetics and Information Technologies, 13(1), 61-72.

Kalbfleisch, J. & Prentice, R. (2002). Competing risks and multistate models [Journal Article]. The Statistical Analysis of Failure Time Data, 247, 277.

Kanakana, G. & Olanrewaju, A. (2011). Predicting student performance in engineering education using an artificial neural network at tswhane university of technology [Conference Proceedings]. ISEM.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations [Journal Article]. Journal of the American Statistics Association, 53(282), 457-481.

Keiding, N., Andersen, P. K. & Klein, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. Statistics in Medicine, 16(2), 215–224.

Kerby, M. B. (2015). Toward a new predictive model of student retention in higher education: An application of classical sociological theory. Journal of College Student Retention: Research, Theory and Practice, 17(2), 138–161.

Kiefer, N. M. (1988). Economic duration data and hazard functions [Journal Article]. Journal of Economic Literature, 26(2), 646-679.

Kim, S. (2014). The comparison of discrete and continuous survival analysis (Thesis).

Kimber, J., Copeland, L., Hickman, M., Macleod, J., McKenzie, J., De Angelis, D. & Robertson, J. R. (2010). Survival and cessation in injecting drug users: prospective observational study of outcomes and effect of opiate substitution treatment [Journal Article]. Bmj, 341, c3172.

Kirby, N. F. (2013). Exploring foundation life science student performance: potential for remediation? (Unpublished doctoral dissertation). University of KwaZulu Natal.

Kirby, N. F. & Dempster, E. R. (2014). Using decision tree analysis to understand foundation science student performance. insight gained at one south african university [Journal Article]. International Journal of Science Education, 36(17), 2825-2847.

Klein, J. P. & Moeschberger, M. L. (2006). Survival analysis: techniques for censored and truncated data [Book]. Springer Science and Business Media.

Kotsiantis, S. B., Pierrakeas, C. & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques [Conference Proceedings]. In International conference on knowledge-based and intelligent information and engineering systems (p. 267-274). Springer.

Kraak, A. (2008). The education-economy relationship in south africa, 2001-2005. Human Resources Development Review, 28, 1–25.

Lancaster, T. (1979). Econometric methods for the duration of unemployment [Journal Article]. Econometrica: Journal of the Econometric Society, 939-956.

Lancaster, T. (1985). Generalised residuals and heterogeneous duration models: With applications to the weilbull model. Journal of Econometrics, 28(1), 155–169.

Lassibille, G. & Navarro Gómez, L. (2008). Why do higher education students drop out? evidence from spain [Journal Article]. Education Economics, 16(1), 89-105.

Lee, E. T. & Go, O. T. (1997). Survival analysis in public health research [Journal Article]. Annual Review of Public Health, 18(1), 105-134.

Lee, S.-H., Lee, K.-D. & Kim, Y. C. (2017). The effects of interest rates, stock prices and trading day to the duration of daily exchange rate pattern: using survival analysis [Journal Article]. International Journal of Monetary Economics and Finance, 10(3-4), 404-425.

Lesik, S. A. (2007). Do developmental mathematics programs have a causal impact on student retention? an application of deicrete-time survival and regression-discontinuity analysis [Journal Article]. Research in Higher Education, 48(5), 583-608.

Letseka, M. (2009). University dropout and researching(lifelong) learning and work (Report). Learning/Work.

Letseka, M. & Breier, M. (2008). Student poverty in higher education: the impact of higher education dropout on poverty. In S. Maile (Ed.), Education and poverty reduction strategies: issues of policy. HSRC Press.

Letseka, M., Cosser, M. & Breier, M. (2010). Student retention and graduate destinations:higher education, labour market access and success. In (p. 90-105). HSRC Press.

Leung, K.-M., Elashoff, R. M. & Afifi, A. A. (1997). Censoring issues in survival analysis [Journal Article]. Annual Review of Public Health, 18(1), 83-104.

Leung, M.-K., Rigby, D. & Young, T. (2003). Entry of foreign banks in the people's republic of china: a survival analysis [Journal Article]. Applied Economics, 35(1), 21-31.

Li, R. (2014). Traffic incident duration analysis and prediction models based on the survival analysis approach [Journal Article]. IET Intelligent Transport Systems, 9(4), 351-358.

Lourens, A. & Smit, I. (2003). Retention: predicting first-year success: research in higher education [Journal Article]. South African Journal of Higher Education, 17(2), 169-176.

Louw, C., Hofman, W. & Van Wyk, B. (2013). Mathematics: A powerful pre-and post-admission variable to predict success in engineering programmes at a university of technology. Perspectives in Education, 31(4), 114–128.

Lunde, A. & Timmermann, A. (2004). Duration dependence in stock prices: An analysis of bull and bear markets [Journal Article]. Journal of Business and Economic Statistics, 22(3), 253-273.

Lunn, P. D. (2010). The sports and exercise life-course: A survival analysis of recall data from ireland [Journal Article]. Social Science and Medicine, 70(5), 711-719.

Mangara, B. (2019). Class attendance and performance of undergraduate electronic engineering students: exploring the effects of gender. In 2019 southern african universities power engineering conference/robotics and mechatronics/pattern recognition association of south africa (saupec/robmech/prasa) (pp. 265–268).

Maree, J. G. (2015). Barriers to access to and success in higher education: Intervention guidelines. South African Journal of Higher Education, 29(1), 390–411.

Maree, J. G., Pretorius, A. & Eiselen, R. J. (2003). Predicting success among first-year engineering students at the rand afrikaans university. Psychological Reports, 93(2), 399–409.

Marimo, M. & Chimedza, C. (2017). Survival analysis of bank loans in the presence of long-term survivors [Journal Article]. South African Statistical Journal, 51(1), 199-216.

Marnewick, C. (2012). The mystery of student selection: are there any selection criteria? Educational Studies, 38(2), 123–137.

Mashiloane, L. & Mchunu, M. (2013). Mining for marks: a comparison of classification algorithms when predicting academic performance to identify students at risk. In Mining intelligence and knowledge exploration (pp. 541–552). Springer.

Meeker, W. Q., Escobar, L. A. & Hong, Y. (2009). Using accelerated life tests results to predict product field reliability [Journal Article]. Technometrics, 51(2), 146-161.

Meyer-Waarden, L. (2007). The effects of loyalty programs on customer lifetime duration and share of wallet [Journal Article]. Journal of Retailing, 83(2), 223-236.

Mills, M. (2011). Introducing survival and event history analysis [Book]. Sage Publications.

Min, Y., Zhang, G., Long, R. A., Anderson, T. J. & Ohland, M. W. (2011). Nonparametric survival analysis of the loss rate of undergraduate engineering studenst [Journal Article]. Journal of Egineering Education, 100(2), 349-373.

Müller, H., Prinsloo, P. & Du Plessis, A. (2007). Validating the profile of a successful first year accounting student [Journal Article]. Meditari Accountancy Research, 15(1), 19-33.

Moeketsi, R. M. & Mgutshini, T. (2014). A comparative time review of recruitment and retention at a university in south africa [Journal Article]. African Journal for Physical Health Education, Recreation and Dance, 20(1), 246-264.

Moeschberger, M. & David, H. (1971). Life tests under competing causes of failure and the theory of competing risks [Journal Article]. Biometrics, 909-933.

Moodley, V. (2014). Quality and inequality in the assessment of visual literacy in grade 12 examination papers across six south african languages. Language Matters, 45(2), 204–223.

Moors, G. & Bernhardt, E. (2009). Splitting up or getting married? competing risk analysis of transitions among cohabiting couples in sweden [Journal Article]. Acta Sociologica, 52(3), 227-247.

Möst, S., Pößnecker, W. & Tutz, G. (2016). Variable selection for discrete competing risks models. Quality & Quantity, 50(4), 1589–1610.

Mueller, D. B., Cao, J., Kongar, E., Altonji, M., Weiner, P.-H. & Graedel, a. T. (2007). Service lifetimes of mineral end uses [Journal Article]. Research Supported by the US Geological Survey, Award(06HQGR0174).

Murphy, T. E., Gaughan, M., Hume, R. & Moore Jr, S. G. (2010). College graduation rates for minority students in a selective technical university: Will participation in a summer bridge program contribute to success? [Journal Article]. Educational Evaluation and Policy Analysis, 32(1), 70-83.

Murray, M. (2014). Factors affecting graduation and student dropout rates at the university of kwazulu-natal [Journal Article]. South African Journal of Science, 110(11-12), 01-06.

Murtaugh, P. A., Burns, L. D. & Schuster, J. (1999). Predicting the retention of university students [Journal Article]. Research in Higher Education, 40(3), 355-371.

Muthen, B. & Masyn, K. (2005). Discrete-time survival mixture analysis [Journal Article]. Journal of Educational and Behavioral Statistics, 30(1), 27-58.

Nam, C. W., Kim, T. S., Park, N. J. & Lee, H. K. (2008). Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies [Journal Article]. Journal of Forecasting, 27(6), 493-506.

Narendranathan, W. & Stewart, M. B. (1993). Modelling the probability of leaving unemployment: competing risks models with flexible base-line hazards [Journal Article]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 42(1), 63-83.

Ndlovu, B. D. (2015). Modelling time to graduation of durban university of technology students using event history analysis. (Unpublished doctoral dissertation).

Neethling, L. (2015). The determinants of academic outcomes: A competing risks approach. In Proceedings of the 2015 conference of the economic society of south africa (pp. 1–17).

Neumann, Y. & Finaly-Neumann, E. (1989). Predicting juniors' and seniors' persistence and attrition: A quality of learning experience approach [Journal Article]. The Journal of Experimental Education, 57(2), 129-140.

Nevo, D. & Ritov, Y. (2013). Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods [Journal Article]. Journal of Quantitative Analysis in Sports, 9(2), 165-177.

Nicholls, G. M., Wolfe, H., Besterfield-Sacre, M. & Shuman, L. J. (2010). Predicting stem degree outcomes based on eighth grade data and standard test scores [Journal Article]. Journal of Engineering Education, 99(3), 209-223.

Nicoletti, C. & Rondinelli, C. (2010). The (mis) specification of discrete duration models with unobserved heterogeneity: a monte carlo study. Journal of Econometrics, 159(1), 1–13.

Nisbet, R., Elder, J. & Miner, G. (2009). Handbook of statistical analysis and data mining applications [Book]. Amsterdam: Academic Press.

Nonnemaker, J. M., Crankshaw, E. C., Shive, D. R., Hussin, A. H. & Farrelly, M. C. (2011). Inhalant use initiation among us adolescents: Evidence from the national survey of parents and youth using discrete-time survival analysis [Journal Article]. Addictive Behaviors, 36(8), 878-881.

Ortiz, E. A. & Dehon, C. (2013). Roads to success in the belgian french community's higher education system: Predictors of dropout and degree completion at the université libre de bruxelles [Journal Article]. Research in Higher Education, 54(6), 693-723.

Pantages, T. J. & Creedon, C. F. (1978). Studies of college attrition: 1950—1975 [Journal Article]. Review of Educational Research, 48(1), 49-101.

Pascarella, E. T. & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. The Journal of Higher Education, 51(1), 60–75.

Paura, L. & Arhipova, I. (2014). Cause analysis of students' dropout rate in higher education study program [Journal Article]. Procedia-Social and Behavioral Sciences, 109, 1282-1286.

Peng, C.-Y. J., Stage, F. K., St John, E. P. & So, T.-S. H. (2002). The use and interpretation of logistic regression in higher education journals:1988-1999 [Journal Article]. Research in Higher Education, 43(3).

Pennington-Cross, A. (2010). The duration of foreclosures in the subprime mortgage market: a competing risks model with mixing [Journal Article]. The Journal of Real Estate Finance and Economics, 40(2), 109-129.

Petersen, I., Louw, J. & Dumont, K. (2009). Adjustment to university and academic performance among disadvantaged students in south africa [Journal Article]. Educational Psychology, 29(1), 99-115.

Pham, H. T., Yang, B.-S. & Nguyen, T. T. (2012). Machine performance degradation assessment and remaining useful life prediction using proportional hazard model and support vector machine [Journal Article]. Mechanical Systems and Signal Processing, 32, 320-330.

Pillay, A. L. & Ngcobo, H. S. (2010). Sources of stress and support among rural-based first-year university students: An exploratory study. South African Journal of Psychology, 40(3), 234-240.

Plank, S. B., DeLuca, S. & Estacion, A. (2008). High school dropout and the role of career and technical education: A survival analysis of surviving high school [Journal Article]. Sociology of Education, 81(4), 345-370.

Pocock, J. (2012). Leaving rates and reasons for leaving in an engineering faculty in south africa: A case study [Journal Article]. South African Journal of Science, 108(3-4), 60-67.

Pohlman, J. T. & Leitner, D. W. (2003). A comparison of ordinary least squares and logistic regression. Ohio Journal of Science, 5, 118-125.

Pretorius, E. (2002). Reading ability and academic performance in south africa: are we fiddling while rome is burning? Language Matters: Studies in the Languages of Southern Africa, 33(1), 169-196.

Price, D. L. & Manatunga, A. K. (2001). Modelling survival data with a cured fraction using frailty models. Statistics in Medicine, 20(9-10), 1515–1527.

Pyke, S. W. & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention [Journal Article]. Canadian Journal of Higher Education, 23(2), 44-64.

Radcliffe, P. M., Huesman Jr, R. L. & Kellogg, J. P. (2006). Modeling the incidence and timing of student attrition: A survival analysis approach to retention analysis [Conference Paper].

Reibnegger, G., Caluba, H., Ithaler, D., Manhal, S., Neges, H. M. & Smolle, J. (2010). Progress of medical students after open admission or admission based on knowledge tests [Journal Article]. Medical Education, 44(2), 205-214.

Restaino, M. (2008). Dropping out of university of salerno: a survival approach (Report). Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Salerno.

Rohr, S. L. (2012). How well does the sat and gpa predict the retention of science, technology, engineering, mathematics, and business students [Journal Article]. Journal of College Student Retention: Research, Theory and Practice, 14(2), 195-208.

Romero, C. & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12–27.

Rooney, C. (2015). Using survival analysis to identify the determinants of academic exclusion and graduation in three faculties at uct (Unpublished doctoral dissertation). University of Cape Town.

Sampson, L. G. (2011). Student persistence in higher education: a study of the challenges and achievements of a group of historically disadvantaged senior students studying at the university of the western cape (Unpublished doctoral dissertation). Stellenbosch: University of Stellenbosch.

Sartorius, K. & Sartorius, B. (2013). The comparative performance of chartered accountancy students in south africa: The impact of historical legacies [Journal Article]. Development Southern Africa, 30(3), 401-416.

Schaap, P. & Luwes, M. (2013). Learning potential and academic literacy tests as predictors of academic performance for engineering students. Acta Academica, 45(3), 181–214.

Scheike, T. H. & Jensen, T. K. (1997). A discrete survival model with random effects: an application to time to pregnancy [Journal Article]. Biometrics, 318-329.

Schoer, V., Ntuli, M., Rankin, N., Sebastiao, C. & Hunt, K. (2010). A blurred signal? the usefulness of national senior certificate (nsc) mathematics marks as predictors of academic performance at university level. Perspectives in Education, 28(2), 9–18.

Schreiber, B. & Yu, D. (2016). Exploring student engagement practices at a south african university: student engagement as reliable predictor of academic performance [Journal Article]. South African Journal of Higher Education, 30(5), 157-175.

Scott, I., Yeld, N. & Hendry, J. (2007). Higher education monitor: A case for improving teaching and learning in south african higher education [Book]. Council on Higher Education Pretoria.

Scott, M. A. & Kennedy, B. B. (2005). Pitfalls in pathways: Some perspectives on competing risks event history analysis in education research [Journal Article]. Journal of Educational and Behavioral Statistics, 30(4), 413-442.

Seabi, J. (2011). Relating learning strategies, self-esteem, intellectual functioning with academic achievement among first-year engineering students [Journal Article]. South African Journal of Psychology, 41(2), 239-249.

Shaw, T. S. (2011). Transitions from cohabitation: A competing risk analysis [Journal Article]. Review of Market Integration, 3(2), 121-159.

Sikhwari, T. (2014). A study of the relationship between motivation, self-concept and academic achievement of students at a university in limpopo province, south africa [Journal Article]. International Journal of Educational Sciences, 6(1), 19-25.

Singer, J. D. & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events [Journal Article]. Journal of Educational Statistics, 18(2), 155-195.

Singer, J. D. & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence [Book]. Oxford University Press.

Siyengo, N. (2015). The educational and psychosocial experiences of first generation students (Unpublished doctoral dissertation). Stellenbosch: Stellenbosch University.

Sommer, M. & Dumont, K. (2011). Psychosocial factors predicting academic performance of students at a historically disadvantaged university [Journal Article]. South African Journal of Psychology, 41(3), 386-395.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. Interchange, 1(1), 64–85.

Spaull, N. & Makaluza, N. (2019). Girls do better: The pro-female gender gap in learning outcomes in south africa 1995–2018. Agenda, 33(4), 11–28.

Stampen, J. O. & Cabrera, A. F. (1986). Exploring the effects of student aid on attrition [Journal Article]. Journal of Student Financial Aid, 16(2), 4.

Steenkamp, H. & Muyengwa, G. (2018). The transition from high school mathematics to engineering mathematics. In Eleventh south african conference on computational and applied mechanics.

Steenkamp, L., Baard, R. & Frick, B. (2009). Factors influencing success in first-year accounting at a south african university: A comparison between lecturers' assumptions and students' perceptions. South African Journal of Accounting Research, 23(1), 113–140.

Stepanova, M. & Thomas, L. (2002). Survival analysis methods for personal loan data [Journal Article]. Operations Research, 50(2), 277-289.

Steyn, T., Owen, R. & Du Preez, J. (2008). Mathematical preparedness for tertiary mathematics-a need for focused intervention in the first year? Perspectives in Education, 26(1), 49–62.

Strydom, F., Mentz, M. & Kuh, G. (2010). Enhancing success in south africa's higher education: Measuring student engagement [Journal Article]. Acta Academica, 42(1), 259-278.

Styron Jr, R. (2010). Student satisfaction and persistence: Factors vital to student retention [Journal Article]. Research in Higher Education Journal, 6, 1.

Sutherland, T. (2018). Retaining female engineering students at the vaal university of technology. In Proceedings of "11th icebe and 7th icie peesa iii international conference on engineering and business education, innovation and entrepreneurship, and capacity building in higher education" (p. 99).

Sy, J. P. & Taylor, J. M. (2000). Estimation in a cox proportional hazards cure model. Biometrics, 56(1), 227–236.

Tan, M. & Shao, P. (2015). Prediction of student dropout in e-learning program through the use of machine learning method [Journal Article]. International Journal of Emerging Technologies in Learning (iJET), 10(1), 11-17.

Therneau, T. M., Grambsch, P. M. & Fleming, T. R. (1990). Martingale-based residuals for survival models [Journal Article]. Biometrika, 77(1), 147-160.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. Review of Educational Research, 45(1), 89–125.

Tinto, V. (1993). Building community. [Journal Article]. Liberal Education, 79(4), 16–21.

Tinto, V. (2010). From theory to action: Exploring the institutional conditions for student retention [Book Section]. In Higher education: Handbook of theory and research (p. 51-89). Springer.

Trussell, J. & Richards, T. (1985). Correcting for unmeasured heterogeneity in hazard models using the heckman-singer procedure [Journal Article]. Sociological Methodology, 15, 242-276.

Tutz, G. (1995). Competing risks models in discrete time with nominal or ordinal categories of response [Journal Article]. Quality and Quantity, 29(4), 405-420. Retrieved from https://doi.org/10.1007/BF01106065 doi: 10.1007/bf01106065

Tutz, G. & Schmid, M. (2016). Modeling discrete time-to-event data [Book]. Springer.

Ullah, S., Gabbett, T. J. & Finch, C. F. (2014). Statistical modelling for recurrent events: an application to sports injuries [Journal Article]. Br J Sports Med, 48(17), 1287-1293.

Uysal, F. (2012). The mathematics education for the engineering students of 21st century. The Online Journal of New Horizons in Education, 2(2), 65–72.

van Broekhuizen, H. & Spaull, N. (2017). The 'martha effect': The compounding female advantage in south african higher education (Report). Stellenbosch University, Matieland South Africa.

van Broekhuizen, H., van der Berg, S. & Hofmeyr, H. (2016). Higher education access and outcomes for the 2008 national matric cohort (Tech. Rep.). Stellenbosch University, Matieland South Africa: Stellenbosch University, Department of Economics.

Vance, C. & Geoghegan, J. (2002). Temporal and spatial modelling of tropical deforestation: a survival analysis linking satellite and household survey data [Journal Article]. Agricultural Economics, 27(3), 317-332.

Van den Berg, G. J. (2001). Duration models: specification, identification and multiple durations [Book Section]. In Handbook of econometrics (Vol. 5, p. 3381-3460). Elsevier.

Van Der Haert, M., Arias Ortiz, E., Emplit, P., Halloin, V. & Dehon, C. (2014). Are dropout and degree completion in doctoral study significantly dependent on type of financial support and field of research? Studies in Higher Education, 39(10), 1885–1909.

Van der Merwe, D. & De Beer, M. (2006). Challenges of student selection: Predicting academic performance. South African Journal of Higher Education, 20(4), 547–562.

Van Praag, C. M. (2003). Business survival and success of young small business owners [Journal Article]. Small Business Economics, 21(1), 1-17.

Van Rooy, B. & Coetzee-Van Rooy, S. (2015). The language issue and academic performance at a south african university [Journal Article]. Southern African Linguistics and Applied Language Studies, 33(1), 31-46.

Van Zyl, L. E. & Rothmann, S. (2012). Flourishing of students in a tertiary education institution in south africa. Journal of Psychology in Africa, 22(4), 593–599.

Vaupel, J. W., Manton, K. G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality [Journal Article]. Demography, 16(3), 439-454.

vd Walt, M. C. & Naidu, M. Y. (2011). Student dropout at the vaal university of technology: A case study (Report). Vaal University of Technology, Vanderbijlpark.

Venkatas, I., Rampersad, N. & Mashige, K. (2014). Do national senior certificate results predict first-year optometry students' academic performance at university? South African Journal of Higher Education, 28(2), 550–563.

Vilakazi, H. W. (2002). African indigenous knowledge and development policy. Indilinga African Journal of Indigenous Knowledge Systems, 1(1), 1–5.

Visser, A. & Hanslo, M. (2005). Approaches to predictive studies: Possibilities and challenges. South African Journal of Higher Education, 19(6), 1160–1176.

Wawrzynski, M. R., Heck, A. M. & Remley, C. T. (2012). Student engagement in south african higher education. Journal of College Student Development, 53(1), 106–123.

Wedekind, V. (2013). Nsc pass requirements (Tech. Rep.). Pretoria,South Africa: Umalusi.

Weybright, E. H., Caldwell, L. L., Xie, H., Wegner, L. & Smith, E. A. (2017). Predicting secondary school dropout among south african adolescents: A survival analysis approach. South African Journal of Education, 37(2).

Wienke, A. (2010). Frailty models in survival analysis [Book]. Chapman and Hall/CRC.

Willett, J. B. & Singer, J. D. (1988). Doing data analysis with proportional hazards models: Model building, interpretation and diagnosis. [Conference Paper].

Willett, J. B. & Singer, J. D. (1991). From whether to when: New methods for studying student droput and teacher attrition [Journal Article]. Review of Educational Research, 61(4), 407-450.

Wohlgemuth, D., Whalen, D., Sullivan, J., Nading, C., Shelley, M. & Wang, Y. (2007). Financial, academic, and environmental influences on the retention and graduation of students [Journal Article]. Journal of College Student Retention: Research, Theory and Practice, 8(4), 457-475.

Xie, H., McHugo, G., Drake, R. & Sengupta, A. (2003). Using discrete-time survival analysis to examine patterns of remission from substance use disorder among persons with severe mental illness [Journal Article]. Mental Health Services Research, 5(1), 55-64.

Yamaguchi, K. (1991). Event history analysis [Book]. California: Sage.

Yang, D., Sinha, T., Adamson, D. & Rosé, C. P. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In Proceedings of the 2013 nips data-driven education workshop (Vol. 11, p. 14).

Yoon, H. & Currid-Halkett, E. (2015). Industrial gentrification in west chelsea, new york: Who survived and who did not? empirical evidence from discrete-time survival analysis [Journal Article]. Urban Studies, 52(1), 20-49.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A. & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year [Journal Article]. Journal of Data Science, 8(2), 307-325.

Zewotir, T., North, D. & Murray, M. (2011). Student success in entry level modules at the university of kwazulu-natal [Journal Article]. South African Journal of Higher Education, 25(6), 1233-1244.

Zewotir, T., North, D. & Murray, M. (2015). The time to degree or dropout amongst full-time master's students at university of kwazulu-natal. South African Journal of Science, 111(9-10), 01–06.

Zhang, T. (2003). A monte carlo study on non-parametric estimation of duration models with unobserved heterogeneity (Tech. Rep.). Department of Economics, University of Oslo.

Zulu, C. (2008). An exploratory study of first-year students at a historically black university campus in south africa: Their academic experiences, success and failure. Africa Education Review, 5(1), 30–47.